Determining and Utilizing the Quasispecies of the Hepatitis B Virus in Clinical Applications

Author
Bastian Beggel

Dissertation

for obtaining the degree
of a Doctor of the Natural Sciences (Dr. rer. nat.)
of the Natural-technical Faculty I
of the Saarland University

Saarbrücken
April, 2014

Bestimmung und Verwendung der Quasispecies des Hepatitis-B-Virus in Klinischen Anwendungen

Autor
Bastian Beggel

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften (Dr. rer. nat.)
der Naturwissenschaftlich–Technischen Fakultät I
der Universität des Saarlandes

Saarbrücken
April, 2014

Tag des Kolloquiums:	08.07.2014
Dekan:	Prof. Dr. Markus Bläser
Vorsitzender des Prüfungsausschusses:	Prof. Dr. Bernt Schiele
Erstgutachter:	Prof. Dr. Thomas Lengauer, Ph.D.
Zweitgutachter:	Dr. Marcel Schulz
Beisitzer:	Dr. Nico Pfeifer

Abstract

Chronic hepatitis B caused by infection with the hepatitis B virus (HBV) affects about 240 million people worldwide and is one of the major causes of severe liver cirrhosis and liver cancer. Hepatitis B treatment options have improved dramatically in the last decade. Effective direct-acting antiviral drugs, so-called nucleos(t)ide analogs, and one effective immunomodulatory drug (pegylated interferon α-2a) are available presently. Current challenges for treating HBV involve the careful selection of patients who require therapy and the thoughtful choice of the treatment option tailored to each patient individually. Personalized medicine aims to optimize treatment decisions based on the analysis of host factors and virus characteristics.

The population of viruses within a host is called the viral quasispecies. This thesis provides statistical methods to infer relevant information about the viral quasispecies of HBV to support treatment decisions. We introduce a new genotyping methodology to identify dual infections, which can help to quantify the risk of interferon therapy failure. We present a method to infer short-range linkage information from Sanger sequencing chromatograms, a method to support treatment adjustment after the development of resistance to nucleos(t)ide analogs. Additionally, we provide the first full-genome analysis of the G-to-A hypermutation patterns of the HBV genome. Hypermutated viral genomes form a subpopulation of the quasispecies caused by proteins of the human innate immune system editing the genome of exogenous viral agents. We show that hypermutation is associated with the natural progression of hepatitis B, but does not correlate with treatment response to interferon.

Kurzfassung

Die Hepatitis-B-Erkrankung wird durch eine Infektion mit dem Hepatitis-B-Virus (HBV) verursacht. Weltweit sind schätzungsweise 240 Millionen Menschen chronisch infiziert. Dabei stellt Hepatitis-B eine der häufigsten Ursachen für die Entwicklung von Leberzirrhose und Leberkrebs dar. Die Behandlungsmöglichkeiten wurden in den letzten zehn Jahren signifikant verbessert. Mittlerweile stehen effektive direkt antivirale Medikamente – sogenannte Nukleos(t)id-Analoga – und ein effektives immunmodulierendes Medikament (pegyliertes Interferon α-2a) für die Behandlung zur Verfügung. Zentrale Fragen bei der Behandlung von Hepatitis-B beinhalten die zielgerichtete Auswahl der Patienten, welche therapiert werden müssen, sowie die passgenaue Auswahl der Behandlungsoption. Die personalisierte Medizin verfolgt das Ziel, die Behandlung basierend auf der Analyse von Patientencharakteristika und Eigenschaften des Virus zu optimieren.

Die Gesamtheit der Viren innerhalb eines Wirtes wird als virale Quasispezies bezeichnet. Diese Arbeit stellt statistische Methoden zur Verfügung, um relevante Informationen über die Quasispezies von HBV zur Unterstützung von Therapieentscheidungen zu ermitteln. Wir entwickeln eine neue Methode zur Genotypisierung, welche Zweifachinfektionen mit HBV identifiziert und somit hilfreich sein kann, das Risiko eines Therapieversagens einer Interferonbehandlung korrekt einzuschätzen. Des Weiteren stellen wir eine Methode vor, welche Linkage-Informationen der viralen Quasispezies, basierend auf den Chromatogrammen der DNA-Sequenzierung nach Sanger, extrahieren kann. Diese Methode kann bei der Umstellung einer Therapie mit Nukleos(t)id-Analoga nach Resistenzentwicklung verwendet werden. Schließlich präsentieren wir die erste Vollgenomanalyse der G-zu-A Hypermutationsmuster von HBV. Hypermutierte virale Genome stellen eine Teilmenge der Quasispezies dar, welche durch von Proteinen der angeborenen Immunabwehr bewirkte Mutationen im viralen Genom entsteht. Wir zeigen, dass diese Subpopulation mit dem natürlichen Verlauf einer Hepatitis-B-Erkrankung, jedoch nicht mit dem Therapieansprechen auf Interferon, statistisch signifikant assoziiert werden kann.

Acknowledgments

Foremost, I would like to thank my advisor Prof. Thomas Lengauer for his encouragement, support, and advice throughout the years. He allowed me to follow my own research ideas and pinpointed rigorously all of their strengths and weaknesses. Thank you for giving me the opportunity to participate in this great research area. I would also like to thank Dr. Marcel Schulz who kindly agreed to referee this thesis.

Special thanks goes to André Altmann and Alexander Thielen who helped me familiarize myself with bioinformatics, in general, and the field of virology, in particular, when I started my thesis.

Special thanks also goes to Rolf Kaiser and all members of his group at the Institute of Virology, University of Cologne. The cooperation with the Rolf Kaiser group was characterized by encouragement, joyousness, and friendship. Maria Neumann-Fraune was my hepatitis B counterpart at the Institute of Virology. She performed a substantial part of the lab work that was required to explore my ideas on Sanger sequencing data. Without her endurance, goodwill, and expertise this thesis would not have been possible.

I would also like to thank Matthias Döring and Valentin Savenko, two of my Bachelor students, for their work on the dual infection model and for many fun discussions.

Thanks to all the collaboration partners of the G-to-A hypermutation project Prof. Andreas Erhardt (Petrus Hospital, Wuppertal), Prof. Carsten Münk (Heinrich-Heine-University, Düsseldorf), and André Boonstra and Harry L. A. Janssen (both Erasmus MC-University Medical Center, Rotterdam).

I would also like to thank Fabian Müller, Maria Neumann-Fraune, Mathieu Flinders, Nele Scharfenberg, Olga Kalinina, Peter Ebert, Prabhav Kalaghatgi, and Simone Susser for proof-reading parts of this thesis.

This work was carried out in the *Department for Computational Biology and Applied Algorithmics* at the *Max Planck Institute for Informatics* in Saarbrücken. I would like to thank all former and current group members. I had five mostly fun and educational but sometimes also painful and exhausting years.

Finally, I would like to thank my parents and especially my mother. You are the best mother that I can imagine. I look up to your friendliness, generosity, and love that you show every day to everybody. I have no idea how you do that.

Contents

1 Introduction

The theory of probabilities is nothing but good sense reduced to calculation; it allows one to appreciate with exactness what accurate minds feel by a sort of instinct, without often being able to explain it.

(Pierre Laplace, 1814)

1.1 Probabilistic Reasoning

The twenty-first century is the age of data. Vast amounts of data are created every day and everywhere: in the internet, in industry, and in research. Technologies to acquire and store these data were developed but the speed of growth in data volume is increasing and the main question remains: how do we make sense of all these data and turn it into meaningful knowledge? The challenges to be addressed are in essence of statistical and computational nature and involve the development and application of statistical learning and reasoning methods. These methods allow us to infer unobserved quantities of the world based on model assumptions and observable evidence.

Probabilistic reasoning is a natural extension of Boolean logic, in which every statement is either true or false, to situations that involve uncertainty (Jaynes, 2003). Uncertainty is ubiquitous in almost all real-world applications either due to restrictions of various kinds that limit our ability to observe and model the world or due to, according to quantum mechanics, the inherently non-deterministic nature of the world. Probabilistic reasoning does nothing more and nothing less than quantifying how the data at hand changes our state of belief about the unobserved quantities of the world. The concept of uncertainty is made precise by the use of probability theory, which, based on very few principles, formalizes human rational reasoning. The *inevitability of probability* to characterize a quantitative system that coherently describes uncertainties has been discussed by various authors (Cox, 1946; de Finetti, 1974; Good, 1950; Lindley, 1982; Ramsey, 1931; Savage, 1961). Nevertheless, the real strength of probability theory lies in its successful application to relevant problems of our age.

In a nutshell, given data $\mathcal{D} = \{x_1, \ldots, x_n\}$ comprised of a set of observations $x_i \in \mathbb{R}^p$, the goal is to infer which of several alternative models $M_1, \ldots, M_m$ matches the data best. Probabilistic reasoning consists of three steps. First, in the modeling step, each model $M_i,\ i = 1, \ldots, m$ is specified as a hypothesis about how a vector of model parameters $\theta_i \in \mathbb{R}^l$ (the unobserved quantities of the world) have led to the data $\mathcal{D}$. In addition, prior distributions $P(M_i)$ and $P(\theta_i|M_i)$ need to be defined that afford the integration of prior knowledge. The formalized description of the data generation process facilitates the computation of the data likelihood for each model given its parameters, $P(\mathcal{D}|M_i, \theta_i)$. Second, in the conditioning step, the probabilistic machinery is applied to derive the so-

called posterior distribution of the model parameters,

$$P(\theta_i|M_i, \mathcal{D}) = \frac{P(\theta_i|M_i)P(\mathcal{D}|M_i, \theta_i)}{P(\mathcal{D}|M_i)},$$

and the probability of each model given the data,

$$P(M_i|\mathcal{D}) = \frac{P(M_i)P(\mathcal{D}|M_i)}{\sum_{j=1}^{m} P(M_j)P(\mathcal{D}|M_j)},$$

with

$$P(\mathcal{D}|M_i) = \int_{\theta_i} P(\theta_i|M_i)P(\mathcal{D}|M_i, \theta_i)\, d\theta_i.$$

We are interested in models and parameter settings, which, given the data we have observed, seem likely or at least more likely compared to other hypotheses. The "best" model might still not reflect the real world, which usually is by far more complicated than what we can include in our computations, but it might provide useful insights into the real world. To test whether or not we have actually learned something is the essential part of the third step, model evaluation.

1.2 Quasispecies of the Hepatitis B Virus

Chronic hepatitis B, defined as the persistent infection with the hepatitis B virus (HBV), is a very complex disease with a highly variable natural progression. Patients may be in one of four phases that are characterized by distinctive profiles in the activity of the host's immune system and the extent of viral replication. Many individuals infected with HBV will develop only mild symptoms during their lifetime while others will die from severe liver cirrhosis or liver cancer. Therefore, patient care has to be individualized starting with the careful selection of patients who require therapy and the thoughtful choice of the treatment option.

This thesis provides statistical methods to analyze the population of hepatitis B viruses within a host, the so-called *viral quasispecies*, to address the challenges of personalized HBV treatment. Determining the viral quasispecies of HBV is a challenging task as up to 10^{11} copies of the viral genome exist within each milliliter serum of an infected individual. The viral population inside a patient may be highly diverse, and it can only be measured indirectly using DNA sequencing[1] technologies of the first or second generation.

First-generation DNA sequencing methods (Sanger sequencing) provide a very rough aggregation of the viral population into a single DNA sequence (so-called population-based sequencing). This aggregation suffers from two major limitations. First, genetic variants of frequency less than 10% to 20% can not be detected. Second, linkage information between genetic variants present at different positions in the genome is lost. Population-based sequencing of the viral quasispecies provides position-wise mixtures of genetic variants, which need to be decomposed using statistical inference to bring different viral variants to light.

Second-generation sequencing technologies have mainly overcome these two limitations and allow for deeper insights into the viral quasispecies. These technologies can amplify and

[1] DNA sequencing is the process of measuring the order of the four types of nucleotides within a DNA molecule.

read-out single viral variants and may provide hundreds of thousands of DNA sequences for each patient sample. Each of these sequences corresponds to an individual member of the viral population. Thus, linkage information is provided over the full read length, which, as of 2014, is in the range of 500 bases. Depending of the number of sequences that were generated, minor variants with relative frequencies as low as 10^{-4} can be detected. Nevertheless, analyzing the viral quasispecies of HBV using second-generation sequencing technologies in clinical applications has its own set of limitations and challenges, millions of error-prone bases have to be analyzed and reduced to interpretable entities.

1.3 Outline

Chapter 2 provides the biological, medical, and statistical background of the thesis. We start with a brief overview of the epidemiology of hepatitis B, followed by a description of the natural course of the disease. Then, we focus on the causative agent of hepatitis B, the hepatitis B virus, its proteins, its genome, and its replication cycle. We discuss the genetic diversity of HBV and introduce genotypically homogeneous subgroups of HBV, the HBV genotypes. We review current treatment options for HBV and state-of-the-art DNA sequencing technologies. We conclude with an introduction to methods and models from the fields of probabilistic reasoning and statistical learning.

In Chapter 3 we discuss the simultaneous infection with two heterologous HBV strains, referred to as *HBV dual infection*. We introduce *in vitro* methods to identify and genotype dual infections and discuss state-of-the-art *in silico* genotyping methods for HBV. We derive the dual infection model, a probabilistic model to infer dual infections based on population-based sequencing data. Results based on the analysis of synthetic test data and real patient sera are presented and discussed.

In Chapter 4 we continue to discuss the problem of determining the viral quasispecies of HBV based on population-based sequencing data. We show that short-range linkage information can be extracted from sequencing chromatograms using a technical artifact of the Sanger sequencing technology, the effect of sequence context-dependent incorporation of dideoxynucleotides. The model is evaluated on *in vitro* and *in silico* test mixtures. The potential and the limitations of our computational approach are discussed in detail.

In Chapter 5 we present a study that relates the phenomenon of G-to-A hypermutation of the HBV genome to the clinical course of hepatitis B. This Chapter is based on second-generation sequencing data that facilitates the analysis of the viral quasispecies in great detail. We will show that the patterns of hypermutation can be linked to the natural progression of hepatitis B and to the replication cycle of HBV.

In Chapter 6 we try to relate treatment response to interferon with G-to-A hypermutation. We describe a prediction model that achieved high accuracy on our training data but was not successfully validated in a follow-up study.

Chapter 7 concludes the thesis and provides an outlook for further advances in personalized HBV therapy.

2 Background

As soon as we touch the complex processes that go on in a living thing, be it plant or animal, we are at once forced to use the methods of this science [chemistry]. No longer will the microscope, the kymograph, the scalpel avail for the complete solution of the problem. For the further analysis of these phenomena which are in flux and flow, the investigator must associate himself with those who have labored in fields where molecules and atoms, rather than multicellular tissues or even unicellular organisms, are the units of study.

(John Jacob Abel, 1 915)

This Chapter outlines the biological, medical, and methodical background of the thesis. In Section 2.1 a comprehensive overview of the hepatitis B disease and its epidemiology is provided. Then, we focus on the causative agent, the hepatitis B virus, in Section 2.2 and on hepatitis B therapy in Section 2.3. Section 2.4 introduces state-of-the-art DNA sequencing technologies. Last, Section 2.5 presents a brief introduction to probabilistic reasoning and two special classes of statistical models, hidden Markov models and support vector machines.

2.1 Hepatitis B

2.1.1 Disease and Epidemiology

Worldwide an approximate number of 240 million people are chronically infected with the hepatitis B virus with mortality rates of about 600,000 per year (WHO, 2013). Chronic hepatitis B infections are defined by the persistence of the *hepatitis B surface antigen* (HBsAg) in sera for more than six months. Hepatitis B is most prevalent in China and South East Asia where nearly 10% of the adult population is chronically infected and the lifetime risk of being infected with hepatitis B is more than 60%. About 45% of the world's population lives in areas where hepatitis B is highly prevalent (more than 8% of the population is HBsAg-positive), which aside from Asia includes the Amazone region, sub-Saharan Africa, and Saudi Arabia (Figure 2.1). Hepatitis B is also highly prevalent in the native populations of Alaska, Greenland, and Northern Canada. Moreover, 43% of the world's population lives in areas where the prevalence is intermediate (between 2% and 8% of the population is HBsAg-positive) and 12% lives in areas of low endemicity (less than 2% of the population is HBsAg-positive). Despite the low prevalence of about 0.6% of hepatitis B in Germany, it is regarded as a major health problem in Germany, too. In 2011, the Robert Koch Institute ranked 127 pathogens to establish strategic priorities for the German national health surveillance and listed the hepatitis B virus in the top 5 list of national health threats (Balabanova et al., 2011).

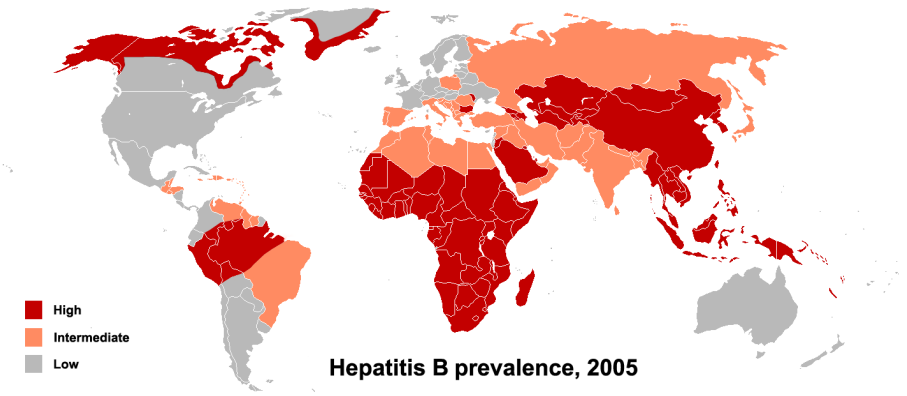

Figure 2.1: Prevalence of hepatitis B as of 2005. Figure obtained from Wikipedia (`http://upload.wikimedia.org/wikipedia/commons/0/09/HBV_prevalence_2005.png`).

Disease progression. During lifetime 15% to 40% of HBsAg carriers develop severe *cirrhosis*[1] or *hepatocellular carcinoma*[2] (HCC) (Hsieh et al., 1992; Lok and McMahon, 2009; Mahoney, 1999). Worldwide approximately 30% of cirrhosis and 53% of HCC can be attributed to hepatitis B infections (Lavanchy, 2004). Liver damage is not caused by HBV directly but develops as a consequence of the host's immune response to the infection. Cirrhosis is divided into two stages: compensated and decompensated. Compensated cirrhosis indicates that the function of the liver is not significantly impaired. Patients with compensated cirrhosis experience few or no symptoms. Decompensated cirrhosis, on the other hand, implies that the scarring of the liver is in a progressed state such that essential body functions are disrupted. Effective long-term suppression of the HBV replication leads to substantial regression of cirrhosis (Marcellin et al., 2013). Otherwise, progression to decompensated cirrhosis is accompanied with diarrhea, abdominal pain, fatigue, and general weakness among other symptoms (Hoofnagle et al., 2007; Lok and McMahon, 2009). As the disease progresses more severe clinical complications are unpreventable, especially ascites (fluid retention in the abdomen), jaundice, internal bleeding, and hepatic encephalopathy. The later arises as the liver loses its main function and is no longer able to remove toxins from the blood serum. Carriers with these complications require liver transplantations to prevent death.

Vaccine. In 1965, Baruch S. Blumberg discovered the hepatitis B surface antigen in the serum of an Australian aborigine while he studied sera of leukemia patients (Blumberg et al., 1965). Blumberg named it the *Australia antigen*. Today the original name is still reflected in the name of hepatitis B's immunodominant antigenic determinant, the so-called "a" determinant (detailed in Section 2.2.1). The work of Blumberg, for which he was honored with the Nobel Prize in Physiology or Medicine, later led to the development of a vaccine against hepatitis B that was available since 1982 (Lemon and Thomas, 1997; Lycke, 1976). This original vaccine contained purified HBsAg particles derived from plasma of chronic hepatitis B carriers. These HBsAg particles were not infectious but could serve as

[1] Cirrhosis is defined as degeneration of liver tissue to scar tissue.
[2] Hepatocellular carcinoma is the most common type of primary liver cancer.

a template to build antibodies against HBsAg (anti-HBs). Since 1989 the production of the modern vaccine uses genetically modified yeast cells, into which the HBsAg gene has been inserted (Lemon and Thomas, 1997; McAleer et al., 1984). Infections with the hepatitis B virus can effectively be prevented by vaccination (Chen, 2009; Margolis, 1993; Viviani et al., 1999; Szmuness et al., 1980). According to the World Health Organization, 179 of their member states routinely vaccinate infants against hepatitis B (WHO, 2013). As of today over one billion doses of the hepatitis B vaccine have been administered worldwide. This has lead to significant reductions in the rates of chronic HBV infections to less than 1% among immunized children in many countries, where more than 8% of children used to become chronically infected (WHO, 2013).

2.1.2 Acute Infections and Chronification

The hepatitis B virus can be transmitted either vertically (perinatally) or horizontally (through contact with blood or other body fluids of an infected person). The first six months of a hepatitis B infection are referred to as the acute phase of the infection. Approximately two thirds of acute hepatitis B infections are clinically asymptomatic (Hoofnagle et al., 2007). Symptoms of acute hepatitis B infections may include jaundice, dark urine, abdominal pain, joint pain, nausea, emesis, and diarrhea. With an incidence of 0.5% to 1% the infection leads to a quick decomposition of the liver referred to as fulminant hepatitis (Lee, 1993).

Approximately 90% of infected adults achieve seroconversion from HBsAg to anti-HBs within six months. This marks the endpoint of an acute infection. Chronic infections, on the other hand, are characterized by the persistence of high levels of HBV DNA, HBsAg, and the hepatitis B e antigen (HBeAg) in sera for more than six months. The chronification rate for children lies between 30% and 90% depending on the age. The chronification rate in immunocompromised populations is about 90% (Hoofnagle et al., 2007; WHO, 2013).

2.1.3 Natural Course of Chronic Infections

Chronic infections with the hepatitis B virus involve complex interactions of the virus, the infected cells, and the host's immune system. Their interplay, which may further be influenced by various external factors, determines the progression of the disease and its severity. The natural course of a chronic hepatitis B infection is described by a four phase model (Hoofnagle et al., 2007; Lok and McMahon, 2009; McMahon, 2009; Yim and Lok, 2006). The *immune-tolerant phase*, the *immune-active phase*, the *inactive carrier phase*, and the *reactivation phase* display different characteristics with respect to viral loads, liver inflammation, alanine aminotransferase[3] (ALT) levels, and very importantly the presence of HBeAg. HBeAg is a non-structural HBV protein that acts as a tolerogen[4]. It downregulates and manipulates innate and adaptive immune responses (Chang et al., 1987; Chen et al., 2004). Not all infections undergo all four phases and multiple transitions between the phases are possible (Figure 2.2). The immune-tolerant phase is usually absent in the case of horizontal transmissions but may last for one to four decades if the infection was

[3] Alanine aminotransferase is an enzyme produced most notably in hepatocytes and its presence in serum is a specific indicator of hepatocellular injury.

[4] A substance (often an antigen) that invokes a specific immune non-responsiveness/tolerance.

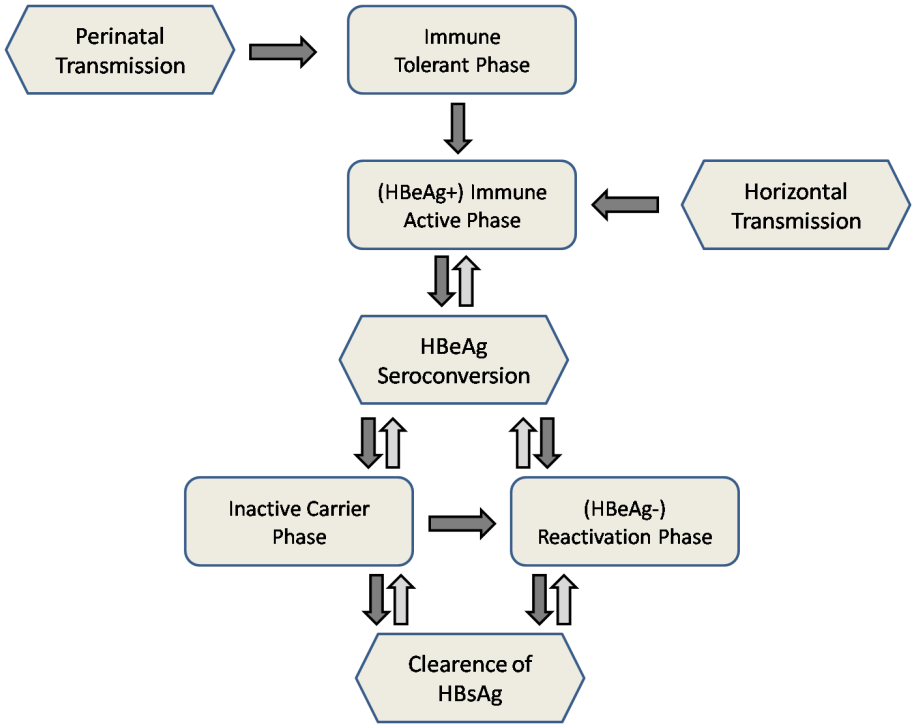

Figure 2.2: Natural course of chronic hepatitis B. Dark arrows indicate typical phase transitions, bright arrows indicate possible reversions of phase transitions. Visualization was adapted from McMahon (2009).

perinatal (Hui et al., 2007). The immune-tolerant phase is characterized by the presence of HBeAg, normal or minimally elevated ALT levels, absent or minimal inflammation, and high levels of HBV DNA (10^8 to 10^{11} copies per milliliter). The prognosis for patients in the immune-tolerant phase is generally good due to the mild inflammation of the liver (Chu et al., 2004). Nevertheless, progression to the immune-active phase occurs when the host's immune system recognizes HBV as an exogenous element. In the immune-active phase liver damage on the cellular level results from the action of the immune system, in particular from cytotoxic T lymphocytes producing antiviral cytokines (Iannacone et al., 2007). The immune-active phase is further characterized by elevated ALT levels, high viral loads (10^6 to 10^{10} copies per milliliter), and the presence of HBeAg. The immune-active phase lasts as long as the host's immune system does not develop antibodies to HBeAg (anti-HBe), referred to as HBeAg seroconversion. HBeAg seroconversion arises spontaneously or treatment induced and is usually followed by the inactive carrier phase, in which ALT levels are normalized, inflammation is absent to minimal, and HBV DNA levels are low or undetectable. Infected persons in the inactive carrier phase control the infection with the presence of anti-HBe antibodies and, thus, again have good prognosis unless the infection is reactivated, which may occur either spontaneously or by immune suppression. In the reactivation phase HBeAg is not expressed (or expressed at low levels) and HBV DNA levels are low to moderate (10^3 to 10^8 copies per milliliter). This phase is further characterized by active liver inflammation indicated by elevated ALT levels. The reactivation phase is also referred to as HBeAg-negative chronic hepatitis. HBV has

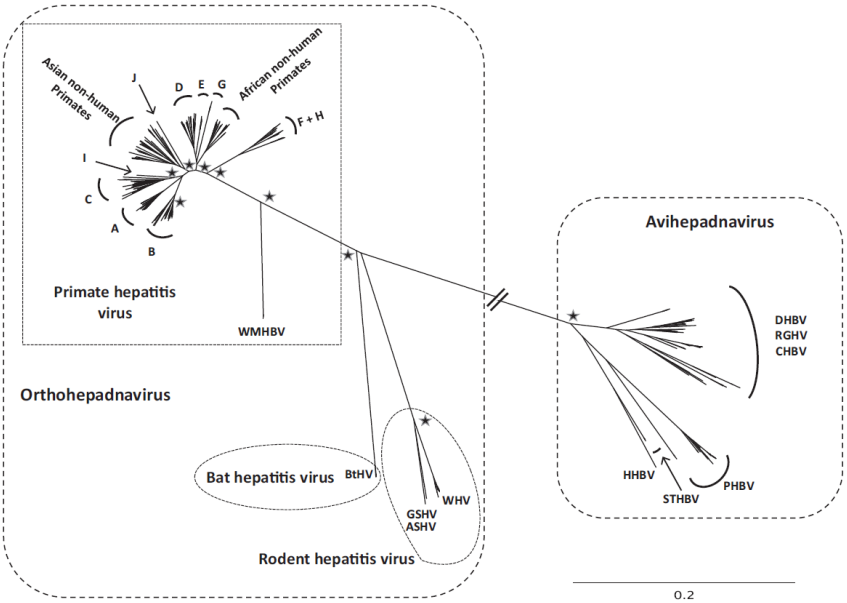

Figure 2.3: The phylogenetic tree shows the genetic relationships between all members of the *Hepadnaviridae* family. 315 complete genomes were used to compute the unrooted neighbor joining tree. The cluster of Asian non-human primates includes gibbons and orangutan hepatitis B virus sequences. The cluster of African non-human primate sequences is composed of chimpanzees and gorilla hepatitis B virus sequences. The stars indicate significant groupings (bootstrap value greater than 85%). Permission to use this visualization was granted by Elsevier (Locarnini et al., 2013).

escaped the anti-HBe immune pressure by specific mutations in the viral genome that suppress or downregulate the expression of HBeAg. The most common of these genetic variations is a single point mutation (G1896A) at nucleotide position 1896, which creates a stop codon in the HBeAg coding region and abolishes the production of HBeAg (Okamoto et al., 1994). Alternatively, a double point mutation (A1762T+G1764A) at positions 1762 and 1764 in the core promoter region downregulates expression of HBeAg (Buckwold et al., 1996; Scaglioni et al., 1997). Often the inactive carrier phase is skipped and patients enter the reactivation phase directly after HBeAg seroconversion. In all phases, HBsAg is present. Seroconversion from HBsAg to anti-HBs, which happens spontaneously at a yearly rate of 0.5% to 1.0%, marks the endpoint of a chronic infection and implies life-long immunity, except for cases of immunosuppression due to chemotherapy or organ transplantation (McMahon et al., 2001).

2.2 Hepatitis B Virus

The hepatitis B virus is a member of the family *Hepadnaviridae*, which is subdivided into two genera, *avihepadnavirus* and *orthohepadnavirus* (Fields et al., 2007). We find *avihepadnaviruses* in birds, e.g. duck hepatitis B virus (DHBV), heron hepatitis B virus (HHBV), crane hepatitis B virus (CHBV), rose goose hepatitis B virus (RGHV), stork

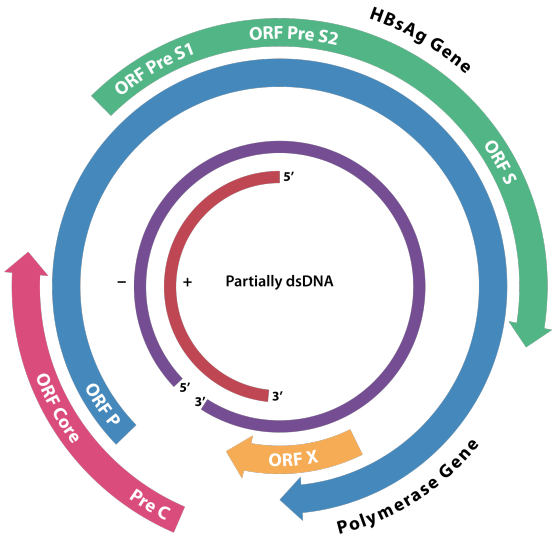

Figure 2.4: Circular representation of the hepatitis B virus genome. Figure was obtained from Wikipedia (`http://upload.wikimedia.org/wikipedia/commons/thumb/0/00/HBV_Genome.svg/2000px-HBV_Genome.svg.png`).

hepatitis B virus (STHBV), and parrot hepatitis B virus (PHBV). *Orthohepadnaviruses* infect mammals, e.g. primates, rodents, and bats. Rodent hepatitis B viruses include, for example, ground squirrel hepatitis B virus (GSHV), artic squirrel hepatitis B virus (ASHV), and woodchuck hepatitis B virus (WHV). Primate hepatitis B viruses include human HBV and several non-human HBV variants. Throughout the thesis HBV always refers to human HBV. Figure 2.3 shows a phylogenetic tree of the *Hepadnaviridae* family. Hepadnaviruses have a very small genome of only 3000 to 3300 bases. The genome has the chemical structure of partially double-stranded, relaxed circular (rc) DNA and duplicates through reverse transcription of an RNA intermediate, the pregenomic (pg) RNA. In general, *hepadnaviruses* have a narrow host range. The woodchuck hepatitis B virus, for example, infects only woodchucks but no other mammals while the ground squirrel hepatitis B virus infects woodchucks and ground squirrels. HBV infects humans and chimpanzees. The woolly monkey hepatitis B virus (WMHBV), on the other hand, is poorly infectious for chimpanzees.

In Section 2.2.1 we discuss the coding elements of the HBV genome and detail the function and structure of the respective proteins. Section 2.2.2 outlines the replication cycle of HBV. In Section 2.2.3 we discuss the genetic diversity of HBV and in Section 2.2.4 we briefly survey the different hypotheses of the origin of hepatitis B in humans.

2.2.1 Particle, Genome, and Proteins

HBV has the smallest genome of all human pathogenic viruses. The genome has been subdivided into distinct HBV genotypes (A to J) based on phylogenetic analysis (see Section 2.2.3). The genomes of all genotypes contain four open reading frames *core*, *surface*, *polymerase*, and *X* (Figure 2.4). Every base in the genome codes for at least one viral protein. About 67% of the genome encode at least two viral proteins. The reading frames

Genomic region	Protein	Reading frame	Positions
Terminal protein (TP)	Polymerase	Polymerase	2307-2843
Spacer (SP)	Polymerase	Polymerase	2844-129
Reverse transcriptase (RT)	Polymerase	Polymerase	130-1161
RNaseH (RN)	Polymerase	Polymerase	1162-1623
PreS1	LHBsAg	Surface	2854-3210
PreS2	LHBsAg, MHBsAg	Surface	3211-154
HBsAg	LHBsAg, MHBsAg, SHBsAg	Surface	155-835
X	X	X	1374-1900
Precore (PC)	HBcAg, HBeAg	Core	1814-1900
Core	HBcAg	Core	1901-2458

Table 2.1: Overview of genomic regions and reading frames in the HBV genome and the proteins these encode for. Positions are given with respect to reference strain AM282986.

are overlapping with frame shifts and all regulatory elements, e.g. promoters, enhancers, and polyadenylation signals, are embedded within coding regions of other genes. The four open reading frames, which correspond to four genes, encode seven viral proteins. Alternative transcriptional start sites facilitate different viral proteins to be generated from a single reading frame. Table 2.1 lists the genomic positions of all relevant coding regions. We refer to these regions throughout the thesis. In the remainder of this Section we detail the function and structure of the viral proteins.

Core gene. The core gene encodes two proteins: the hepatitis B core antigen (HBcAg) and HBeAg. HBcAg is the building block of the HBV capsid. Depending on the HBV genotype it is formed by 183 or 185 amino acids. Nevertheless, HBcAg is highly conserved among the different genotypes likely due to structural requirements of capsid assembly (Chain and Myers, 2005). The N-terminal 149 or 151 amino acids (depending on the genotype) are referred to as the assembly domain that forms the protein shell of the capsid. The remaining amino acids form the so-called protamine domain, which is positively charged due to a high number of arginine residues. The protamine domain binds to HBV's pregenomic RNA and initiates the packing of the viral genome (Gallina et al., 1989; Zlotnick et al., 1997).

Based on cryo-electron microscopy and crystallization studies a detailed picture of the HBV capsid is available (Böttcher et al., 1997; Wynne et al., 1999). The capsid has an icosahedral structure. It consists either of 120 or of 90 HBcAg homodimers (Figure 2.5). The capsid configuration that consists of 120 HBcAg homodimers has a symmetry with triangulation number 4 (T = 4) and is more prevalent in infectious viral particles. The configuration with 90 HBcAg homodimers has a symmetry with T = 3 and a prevalence of about 10% (Crowther et al., 1994; Short et al., 2009). The HBcAg homodimer subunits are paired by forming a four-helix bundle stabilized by an intermolecular disulfide bond (Böttcher et al., 1997; Conway et al., 1997). The HBcAg homodimers form the spikes of the surface of the capsid, which are surrounded by flanking pores. The tips of the spikes display the major immunodominant region on the capsid surface, which is recognized by

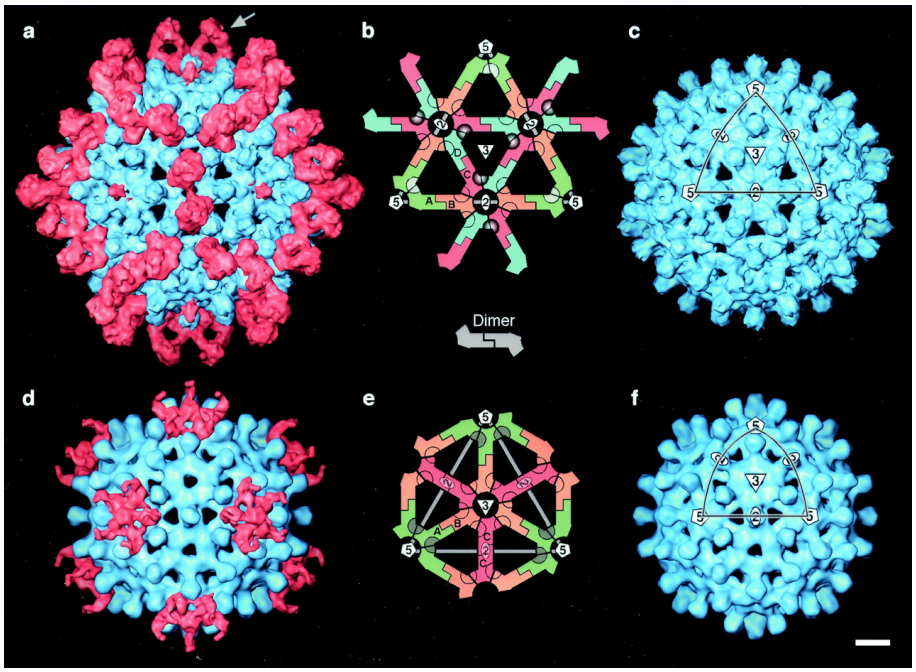

Figure 2.5: Reconstructions of hepatitis B virus capsid based on cryo-electron microscopy. An epitope (Fab-3120) was used for color labeling. Capsid protein is blue, Fab-3120 density is pink. (a) Capsid conformation with triangulation number 4 (T = 4). (d) Capsid conformation with triangulation number 3 (T = 3). (b) and (e) Lattice diagram of single triangular facets of the capsid, as marked on subplots (c) and (f). (c) and (f) Unlabeled capsids with triangulation number 4 and 3, respectively. The bar in the right bottom corner indicates the scale of 50 Å. Figure was obtained from Conway et al. (2003).

the adaptive immune system to develop antibodies (anti-HBc) (Wynne et al., 1999). The antibodies anti-HBc are persistent throughout the infection but do not neutralize HBV. The pores have a diameter between 12 and 15 Å. They allow for diffusion of nucleotides and other small molecules into the interior of the capsid (Böttcher et al., 2006; Perlman et al., 2005).

HBeAg is the second viral protein encoded by the core gene. Its translational start position is 29 codons upstream of HBcAg. HBeAg is non-structural but serves as an important biomarker to monitor the natural progression of hepatitis B as discussed in Section 2.1.3. The function of HBeAg is not completely understood yet (Walsh and Locarnini, 2012). It is not required for viral assembly, infection, or replication (Chang et al., 1987). Nevertheless, it was shown to attenuate the host's immune response to HBcAg and to downregulate the innate and adaptive immune system (Chen et al., 2004; Visvanathan et al., 2007). HBeAg traverses the placenta to induce immune tolerance *in utero*, which promotes viral persistence following perinatal infection (Milich et al., 1990). Thus, HBeAg plays a key role in viral persistence and promotes HBV chronicity due to its immunomodulatory effects.

Surface gene. The envelope consists of a host cell-derived phospholipid bilayer and three viral surface proteins called large (LHBsAg), middle (MHBsAg), and small (SHBsAg) hepatitis B surface antigen (Schädler and Hildt, 2009). Throughout the thesis we use HBsAg to refer to either of the three surface proteins. The three proteins have different start codons but share the same end position in the genome. The coding regions that encode the surface proteins are named preS2, preS1, and S. SHBsAg is encoded by S alone and consists of 226 amino acids. SHBsAg contains the immunodominant antigenic determinant referred to as the "a" determinant. It is located between amino acid positions 124 and 147 of SHBsAg and plays an important role in diagnosis and immunoprophylaxis. Several immune and vaccine escape mutations have been described within the "a" determinant (Carman et al., 1990; Chiou et al., 1997; Lee et al., 2001). MHBsAg has 55 amino acids (encoded by preS1) added to the N-terminus of SHBsAg. PreS2 encodes additional 108, 118, or 119 amino acids (depending on the genotype), which are added to MHBsAg at its N-terminus. The surface proteins are translocated across the endoplasmic reticulum (ER) membrane (Eble et al., 1987, 1990). The structural conformation of the surface proteins with respect to the ER membrane is detailed in the caption of Figure 2.6.

During chronic infection HBsAg is expressed in large amounts (up to 10,000-fold excess over virions) in the form of quasispherical particles and tubes of variable length (Figure 2.7). These empty subviral particles are noninfectious and have a diameter of approximately 20 nanometer (Gilbert et al., 2005). They are formed by SHBsAg and LHBsAg in the lumen of the ER and the ER-Golgi intermediate compartment (Patient et al., 2007). The biological function of these subviral particles has not been fully understood yet. A study performed on duck hepatitis B virus suggests that the subviral particles strongly enhance intracellular viral replication and are required to establish an infection *in vivo* (Bruns et al., 1998).

Polymerase gene. The polymerase gene, which spans almost 80% of the HBV genome, encodes a single protein, the HBV polymerase. The polymerase has four functional domains. The terminal protein (TP) domain initiates capsid packing and reverse transcription by binding to a stem-loop located at the $5'$ of the pregenomic RNA, the so-called ε-signal (Lanford et al., 1997; Wang and Seeger, 1992; Weber et al., 1994). We further detail this in Section 2.2.2. The reverse transcriptase (RT) domain and the RNaseH domain together

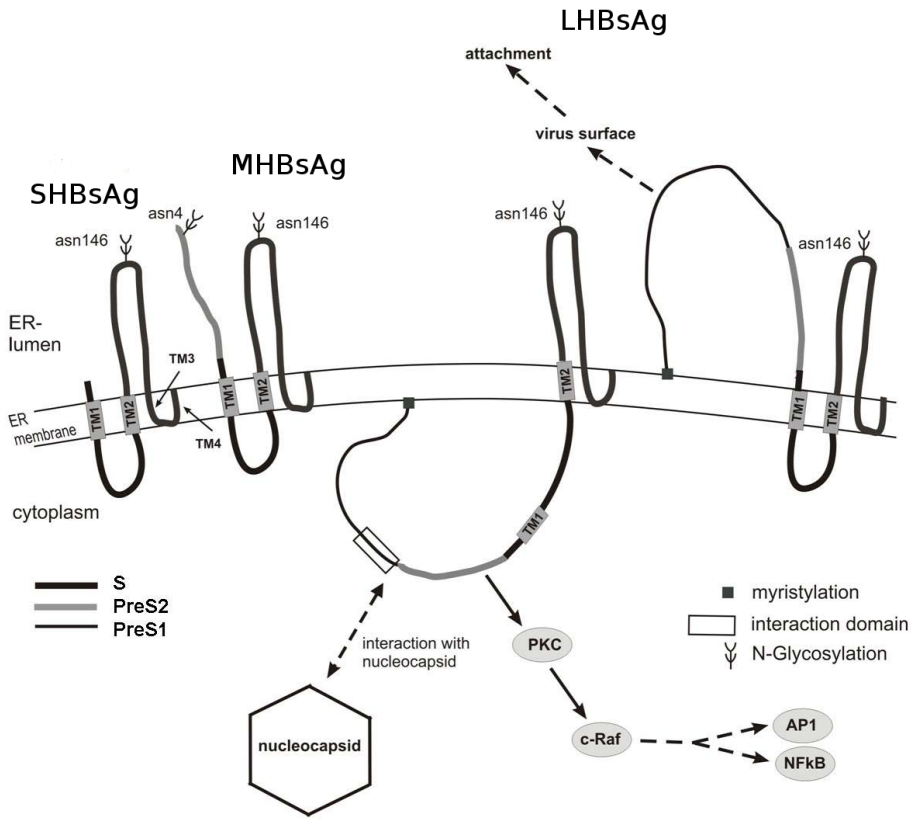

Figure 2.6: Visualization of the three hepatitis B virus surface proteins (SHBsAg, MHBsAg, and LHBsAg) integrated into the endoplasmic reticulum membrane. The first transmembrane region (TM1), located between amino acids 8 and 22 of SHBsAg, initiates the membrane insertion. At amino acid positions 80 to 98 a second transmembrane region (TM2) is located. Position 146 harbors an N-glycosylation site depicted as asn146. SHBsAg contains one or two additional transmembrane regions TM3 and TM4 at its C-terminal region. MHBsAg has 55 amino acids added to the N-terminus of SHBsAg, which are located in the lumen of the endoplasmic reticulum. Aside from this MHBsAg and SHBsAg share the same topology. MHBsAg contains a glycosylation site at amino acid position 4 (asn4). LHBsAg has additional 108, 118, or 119 amino acids (depending on the genotype), which are added to MHBsAg at its N-terminus. LHBsAg has two distinct conformational forms of the N-terminal domain. In one conformation the topology of LHBsAg differs from the topology of SHBsAg and MHBsAg in a way that its N-terminus including TM1 is located inside the cytoplasm. In this conformation LHBsAg establishes the binding to the virus capsid (Bruss, 1997). In the other conformation, which is topologically similar to SHBsAg and MHBsAg, the N-terminus is exposed on the surface of virions and was described to have an important function during cell entry (Prange and Streeck, 1995). Figure obtained from Schädler and Hildt (2009).

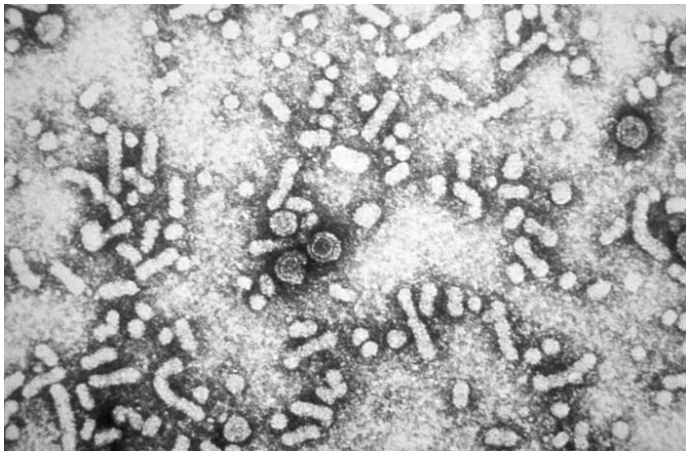

Figure 2.7: Electron microscopic image of infectious HBV virions and subviral particles. Infectious HBV virions have a diameter of about 42 nanometer and contain a genome. Subviral particles, which do not contain a genome and are noninfectious, are either quasispherical particles or tubes of variable length. Subviral particles have a diameter of approximately 20 nanometer and are formed by SHBsAg and LHBsAg. Subviral particles are expressed in large amounts compared to virions. Figure was obtained from Wikipedia (`http://upload.wikimedia.org/wikipedia/commons/1/12/Hepatitis-B_virions.jpg`).

perform the reverse transcription of the pregenomic RNA into partially double-stranded, relaxed circular DNA present in mature particles. The spacer domain was not found to have a specific function other than to provide a spatially flexible connection between the terminal protein domain and the reverse transcriptase domain (Bartenschlager and Schaller, 1988; Chang et al., 1990).

X gene. The X gene encodes the hepatitis B virus X antigen (HBxAg), which has central functions in HBV biology and hepatic pathogenesis by modulating many viral and cellular functions. The X gene is highly conserved among mammalian hepadnaviruses. Its product, HBxAg, can be observed very early after infection and throughout the chronic phase. HBxAg has not been identified in mature virions but can be observed throughout the cytoplasm with accumulation in the perinuclear region. HBxAg consists of 154 amino acids and has two functional domains, which are critical for its regulatory activity (Kumar et al., 1996; Martin-Vilchez et al., 2011; Takada and Koike, 1994). One functional site is at amino acid position 69 and the other is between amino acid positions 110 and 139. HBxAg does not bind to DNA but regulates transcription of cellular genes, including oncogenes, cell growth factors, DNA repair genes, and genes associated with apoptosis (Martin-Vilchez et al., 2011). Several studies investigated protein-protein interactions of HBxAg with human proteins using screening protocols like the yeast two-hybrid assay or immunoprecipitation/mass spectrometry approaches. More than 30 human proteins were reported to interact with HBxAg but the biological significance and mode of action of these virus-host protein interactions are difficult to determine (Fields et al., 2007; Kumar et al., 2011). Studies performed with the woodchuck hepatitis B virus showed that a functional form of HBxAg is required to initiate infections *in vivo* (Chen et al., 1993; Zoulim et al.,

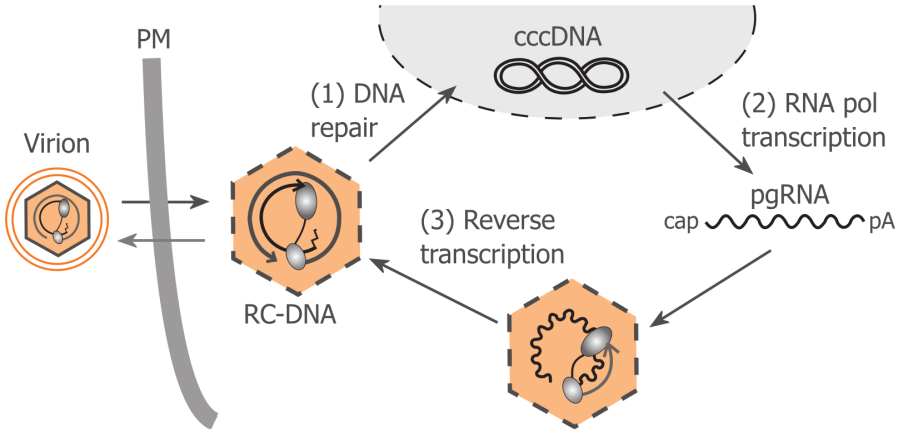

Figure 2.8: Replication cycle of the hepatitis B virus. Cell entry and fusion is followed by transportation of the capsid to the nucleus. (1) After uncoating and uptake of the HBV genome into the nucleus, partially double-stranded, relaxed circular DNA is transformed into covalently closed circular DNA. (2) An RNA intermediate, the pregenomic RNA, is transcribed from the cccDNA. (3) Pregenomic RNA is reverse transcribed to DNA and encapsidated. Newly formed capsids may renter the nucleus or be enveloped by HBV surface proteins in the endoplasmic reticulum and released. Figure obtained from (Beck and Nassal, 2007).

1994).

2.2.2 Replication Cycle

As mentioned before, HBV replicates through reverse transcription of an RNA intermediate. The replication cycle of HBV is summarized in Figure 2.8 and detailed in the remainder of this Section.

Cell entry. The mechanisms of attachment and productive cell entry, as the initial step in the HBV replication cycle, are not fully understood yet (Schädler and Hildt, 2009). According to the general concept of viral infection, cell entry can be divided into three steps: attachment, fusion, and entry. The interaction of HBV's surface proteins with a specific receptor on the host cell membrane mediates the attachment of virions. HBV uses the carbohydrate side chains of hepatocyte-associated heparan sulfate proteoglycans as attachment receptors (Leistner et al., 2008; Schulze et al., 2007). It was shown that this interaction requires the integrity of LHBsAg. The attachment is believed to have low affinity and to be reversible. Thus, it may or may not be followed by cell fusion (Marsh and Helenius, 2006). The complete understanding of the fusion of the viral and the cellular membranes and the subsequent uptake of the virion into the cell is an active research challenge. Lately, the sodium taurocholate cotransporting polypeptide (NTCP) has been identified as a cellular entry receptor for HBV (Ni et al., 2014; Yan et al., 2012).

Integration into the nucleus. The HBV replication cycle requires the HBV genome to be integrated into the nucleus of the host cell. Perinuclear accumulation of the HBV capsid can be observed already about 15 minutes after insertion into the cytoplasm (Schädler and

Hildt, 2009; Sodeik, 2000). The transportation of the genome towards the perinuclear area is a directed process, which utilizes the cellular microtubule system (Rabe et al., 2006).

The uncoating of virions may begin instantly after virus fusion, but this has not been confirmed yet. Likewise, the mechanisms that mediate entry of the viral genome into the nucleus of the host cell are still enigmatic. In mature particles the HBV genome has the chemical structure of partially double-stranded, relaxed circular DNA and has a HBV polymerase covalently attached to the 5′ end of the (–)-DNA strand. In the nucleus the HBV genome persists in the form of covalently closed circular (ccc) DNA. cccDNA is very stable and serves as the central intracellular intermediate during viral replication. As detailed in the paragraph about reverse transcription in this Section, derivation of cccDNA from rcDNA requires the cleavage of the HBV polymerase, the completion of the (+)-DNA strand, the removal of the RNA primer, the cleavage of redundant DNA, and the ligation of the (+)-DNA and (–)-DNA strands.

Transcription and encapsidation. Transcription of three subgenomic[5] RNAs and one supergenomic[6] RNA, the pgRNA, is performed by the cellular RNA polymerase II. Two of the subgenomic RNAs are coding for surface proteins (2.4 kilobase RNA and 2.1 kilobase RNA) and one is coding for HBxAg (0.7 kilobase RNA) (Schädler and Hildt, 2009). The pgRNA has a length of about 3500 bases. Due to its supergenomic length, it contains two copies of the direct repeat 1 element (DR1) and two copies of the ε-signal, a stem-loop structure. It serves as a template for the reverse transcription into rcDNA and as a template for translation of the HBV polymerase, HBcAg, and HBeAg. The TP domain of the polymerase binds preferentially to the very same pgRNA molecule from which it was translated and together they form the pgRNA-polymerase complex (Bartenschlager et al., 1990). This binding at the 5′ proximal ε-signal initiates encapsidation of the pgRNA and its reverse transcription (Nassal, 2008) (Figure 2.9). This mechanism ensures that a single pgRNA molecule and a single polymerase are packaged in each viral particle. The pgRNA-polymerase complex acts as an anchor for the attachment of HBcAg homodimers that leads to capsid formation. At a certain rate newly formed capsids reside inside the cell and reenter the nucleus to increase the number of cccDNA copies available in the cell.

Reverse transcription. The reverse transcription of the pgRNA is a very complex yet well understood process. In short, first the (–)-DNA strand is formed using the pgRNA as a template followed by the synthesis of the (+)-DNA strand resulting in rcDNA. The following outline of the reverse transcription takes reference to Nassal (2008). In detail, DNA synthesis is initiated by protein-priming (Wang and Seeger, 1992). The HBV polymerase uses a bulged region within the 5′ proximal ε-signal stem-loop as a template to generate a four nucleotide long DNA oligonucleotide, which is then translocated together with the covalently attached TP domain to the 3′ proximal DR1 (Figure 2.9). Starting at 3′ proximal DR1 the RT domain synthesizes the (–)-DNA strand until the very 5′ end of the pgRNA. Simultaneously, the pgRNA template is degraded by the RNaseH domain but not until the very end. A small RNA oligomer (11 to 16 nucleotides of the pgRNA), which includes the 5′ proximal DR1 remains (Haines and Loeb, 2007). Next, translocation of the RNA oligomer to DR2 (which is sequence identical to DR1) creates the primer for

[5]Subgenomic RNA refers to an RNA molecule that is shorter than the HBV genome.

[6]Supergenomic RNA refers to an RNA molecule that is longer than the HBV genome. Its transcription is possible due to the circular structure of the HBV genome.

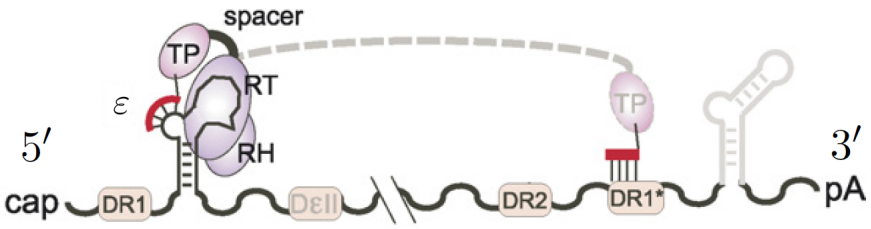

Figure 2.9: The pregenomic RNA contains two copies of the ε-signal, a stem loop, and DR1 (pictured as DR1 and DR1*). The TP domain of the HBV polymerase binds to the 5′ end proximal ε-signal and uses its bulge region to create a four-nucleotide long DNA primer. This DNA oligonucleotide is translocated to the 3′ proximal DR1 element (depicted as DR1*), which initiates the reverse transcription of the pgRNA into the (−)-DNA strand. Permission to use this figure was granted by Elsevier (Nassal, 2008).

the (+)-DNA strand synthesis. With low frequency the RNA oligomer does not switch to DR2, in which case (+)-DNA strand synthesis is initiated from DR1. This alternative route is called *in situ* priming and results in double-stranded linear DNA, which likely is nonreplicative. Nevertheless, translocation and priming from DR2 is predominant. Folding of the (−)-DNA strand facilitates the template switch as it ensures that DR1 and DR2 lie close together in space (Lewellyn and Loeb, 2007). (+)-DNA strand synthesis is facilitated by the RT domain starting from DR2 until the very 5′ end of the (−)-DNA strand. Then, a third translocation of the growing end of (+)-DNA strand mediated by small terminal repeat elements yields circulation and upon further extension of the (+)-DNA stand results in rcDNA. The intermediate steps of the reverse transcription of the pgRNA are shown in detail in Figure 2.10.

Capsid maturation, envelopment, and secretion. As pointed out before, mature HBV particles predominantly contain partially double-stranded, relaxed circular DNA. But HBV DNA found in the cytoplasm of infected cells contain equal amounts of partially double-stranded and single-stranded DNA (Summers and Mason, 1982; Wei et al., 1996). Thus, the synthesis of the (+)-DNA strand likely triggers modifications on the surface of the capsid, which facilitate the interaction with HBV's surface proteins and afford envelopment. The molecular modifications are suspected to involve changes in the phosphorylation state of HBcAg (Liao and Ou, 1995). Translocation of the capsid across the ER membrane is mediated by their interaction with LHBsAg and results in enveloped particles, which are secreted into the bloodstream (Huovila et al., 1992).

2.2.3 Genetic Diversity and HBV Genotypes

The HBV genome has been classified into eight well characterized genotypes A to H based on a nucleotide variation threshold of 8% over the entire genome (Arauz-Ruiz et al., 2002; Naumann et al., 1993; Norder et al., 1994, 2004; Okamoto et al., 1988; Stuyver et al., 2000) and two additional (very rare) genotypes, namely I (Olinger et al., 2008; Tran et al., 2008; Yu et al., 2010) and J (Tatematsu et al., 2009). HBV genotypes have been found

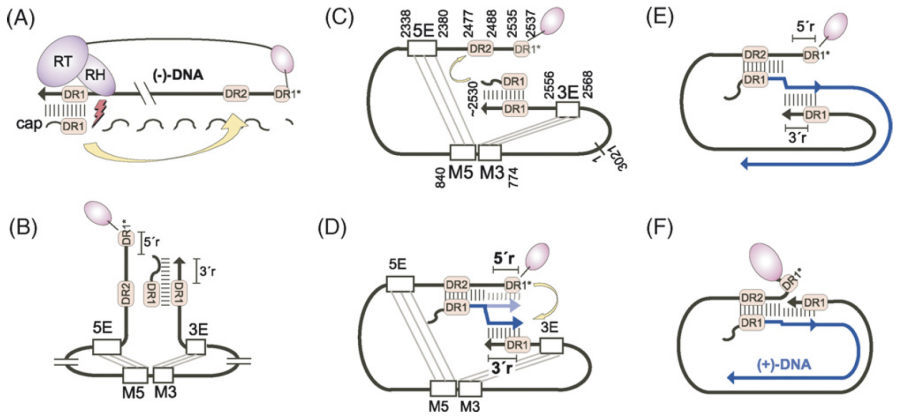

Figure 2.10: Reverse transcription of the pregenomic RNA. (A) The pgRNA is reverse transcribed into (–)-DNA until the very 5′ end. The RNaseH domain of the polymerase digests the pgRNA but leaves a short RNA oligomer, which after translocation to DR2 serves as a primer for the synthesis of the (+)-DNA strand. (B and C) Translocation of the RNA oligomer is supported by bending of the (–)-DNA strand facilitated by functional element 5E, M5, 3E, and M3. (D) The growing end of the (+)-DNA strand is translocated to the 3′ end of the (–)-DNA strand, which creates the relaxed circular structure of the HBV genome. The translocation of the growing end is possible due to small repeat elements 5′r and 3′r. (E and F) Structural representation of partially double-stranded, relaxed circular DNA, the form of the HBV genome present in mature particles. Permission to use this figure was granted by Elsevier (Nassal, 2008).

Genotype	Genome Length (NT)	Insertions and Deletions
A	3221	Insertion of two amino acids at position 153 and 154 in HBcAg
B	3215	
C	3215	
D	3182	Deletion of amino acid 1 to 11 in preS1
E	3212	Deletion of amino acid 11 in preS1
F	3215	
G	3248	Insertion of 36 nucleotides after the 5th base in HBcAg. Deletion of amino acid 11 in preS1
H	3215	
I	3215	
J	3182	Deletion of amino acid 1 to 11 in preS1

Table 2.2: Hepatitis B virus genotypes and genome lengths. The typical genomes of genotypes B, C, F, H, and I have a length of 3215 nucleotides. The typical genomes of genotypes A, D, E, G, and J differ due to genotype specific insertions or deletions. Summary is based on Schaefer (2007); Tatematsu et al. (2009); Tran et al. (2008).

to influence the rate of chronicity (Ogawa et al., 2002; Suzuki et al., 2005; Zhang et al., 2008), the course of the disease (Chan et al., 2004; Kao et al., 2000; Livingston et al., 2007; Sánchez-Tapias et al., 2002; Thakur et al., 2002; Yuen et al., 2004), and the response to interferon treatment (Section 2.3.1). HBV genomes of genotypes B, C, F, H, and I regularly have a length of 3215 nucleotides. The genome of other genotypes differ slightly due to genotype specific insertions or deletions (Table 2.2).

HBV genotypes are further classified into subgenotypes, which are named according to the genotype with a numeric appendix. The following outline takes reference to Lin and Kao (2011). HBV subgenotypes are defined based on a 4% divergence across the complete genome. Genotype A is divided into A1, A2, and A3, which are prevalent in distinct geographic areas, namely sub-Saharan Africa, Northern Europe, and Western Africa, respectively. Genotypes B and C are the dominant genotypes in Asia. B1 is common in Japan and B6 is common in indigenous populations in Alaska, Northern Canada, and Greenland. Subgenotypes B2 to B5, which are prevalent throughout Asia except Japan, have a recombinant core gene that partly consists of genotype C. Subgenotypes C1 to C3 are most prevalent in Taiwan, China, Korea, and Southeast Asia while subgenotype C4 circulates in Australia. C5 is mainly found in the Philippines and in Vietnam. The subgenotypes of genotype D (D1 to D5) are widely spread across Eastern Europe, the Mediterranean region, northern Africa, Russia, the Middle East, and India. To date, no subgenotypes are defined for genotype E, which is restricted to West Africa. Genotypes F and H are phylogenetically closely related. Genotype F is divided into subgenotypes F1 to F4 that are prevalent in Central and South America while genotype H is primarily found in Mexico and Nicaragua (Devesa and Pujol, 2007). Genotype G is the most uncommon genotype among the established genotypes (A to H). Isolates were found in France, Germany, and the United States. Genotype G occurs in persons infected with a second genotype, usually

genotype A. All genotype G genomes harbor the G1896A stop codon mutation such that the second genotype is required to express HBeAg (Kato et al., 2002). The newly identified genotype I was found in Vietnam and Laos. Genotype J was found in one individual that lived in the Ryukyu Islands in Japan.

HBV dual infections, defined as the simultaneous infection with two heterologous HBV strains, may lead to the exchange of genetic material between the viral strains. The resulting hybrids are referred to as recombinant forms or recombinants. Recombination events may arise at multiple steps during the replication cycle of HBV. Frequent inter- and intramolecular recombinations were described at RNA and DNA level during either transcription of cccDNA or reverse transcription of pgRNA (Newbold et al., 1995; Yang and Summers, 1995). Due to their high prevalence in certain geographic regions recombinant forms play an important role for the genetic diversity of HBV. Two distinct versions of C/D recombinants developed into the two most prevalent HBV variants in Tibet (Cui et al., 2002; Wang et al., 2005a). Subgenotypes B2 to B5, which are prevalent throughout Asia except for Japan, have a recombinant core gene that partly consists of genotype C, as mentioned before. A/D recombinants circulate in Italy, South Africa, and India (Chauhan et al., 2008; Morozov et al., 2000; Owiredu et al., 2001). Two studies analyzed full genome HBV sequence data and identified eight genomic regions, in which recombination breakpoints accumulate (Simmonds and Midgley, 2005; Yang et al., 2006).

2.2.4 Origin of Hepatitis B in Humans

The origin of HBV in humans has not yet been resolved (Locarnini et al., 2013). Several models were proposed but none of these are conclusive and widely accepted.

The so-called “Out of South America” hypothesis proposes that HBV originated in the New World and spread about 400 years ago in concert with the colonization (Bollyky et al., 1997). Two lines of arguments contradict this hypothesis. First, this model implies that genotypes F and H (today common in South America) migrated to Africa and Asia to develop a large chronic carrier population and diverse into non-F/H genotypes in only 400 years. Second, HBV is widely distributed in wild living non-human primate species such as chimpanzees and gorillas as well as in gibbons and orangutans (Locarnini et al., 2013). Following the “Out of South America” hypothesis these infections in primates were caused by humans. Both implications seem very unlikely.

HBV may have co-evolved with anatomically modern humans as they migrated out of Africa 100,000 years ago (Norder et al., 1994). As a consequence the genetic relationships between human subpopulations should be reflected in the genetic diversity of HBV endemic in the respective human subpopulations. But this is not the case. For example, the native American population (which is associated with HBV genotype F) is genetically closely related to “Mongolian” North East Asians associated with genotypes B and C. But genotypes F is not closely related to genotypes B and C, which contradicts the “Out of Africa” hypothesis (Locarnini et al., 2013).

Another model proposes several independent transmission events from primates to humans with several distinct genetic variants (Simmonds, 2001). This hypothesis is similar to that on the origins of the human immunodeficiency virus (HIV). Phylogenetic studies identified multiple cross-species transmission events of African non-human primate lentiviruses

(simian immunodeficiency virus (SIV)) as the origin of HIV. HIV-2 likely emerged from sooty mangabey (*Cercocebus atys*) SIV (Hirsch et al., 1989; Gao et al., 1992). Similarly, several subgroups of HIV-1 can be linked to at least three transmission events of chimpanzee SIV from an African chimpanzee species to humans (Gao et al., 1999). The cross-species transmission hypothesis for HBV is supported by the fact that geographic areas of high HBV prevalence (South America, sub-Saharan Africa, and South East Asia) are those in which contact of humans with primates are most likely. Nevertheless, the HBV genotypes present in the respective areas are not particularly closely related to the genomes of the non-human primate viruses in these regions. Thus, the species involved in the transmission(s) to humans are still unidentified and likely have become extinct, which makes the multiple transmissions hypothesis difficult to validate (Locarnini et al., 2013).

2.3 Hepatitis B Therapy

The primary goal of hepatitis B therapy is to prevent the progression of liver degeneration in form of liver cirrhosis and HCC. The main viral factors associated with disease progression are persistent presence of HBeAg and persistently high levels of HBV DNA (Chen et al., 2006; Iloeje et al., 2006; Kwon and Lok, 2011). Thus, long-term suppression of viral replication and possible treatment induced HBeAg seroconversion have to be achieved. The level of HBV DNA in serum is the most important indicator of HBV treatment success or failure. Additionally, normalization of alanine aminotransferase levels, decrease of liver inflammation, and loss of HBeAg are used in clinical routine to measure treatment outcome. The ultimate goal of therapy is the clearance of HBsAg and the seroconversion to anti-HBs. This marks the endpoint of an infection and normally ensures life-long immunity. As of 2014, seven drugs have been approved by the United States Food and Drug Administration for the treatment of hepatitis B: two variants of *interferon* (IFN) and five *nucleos(t)ide analogues* (NA).

2.3.1 Treatment with Interferon

Interferons are glycoproteins that have strong immunomodulatory, antiviral, and antitumor activity. Interferons were first described by Isaacs and Lindenmann (1957) who demonstrated their antiviral activity and named them due to their ability to "interfere" with viral replication. Interferons are signaling proteins that upregulate the transcription of hundreds of so-called interferon-stimulated genes, whose protein products mediate immunomodulatory, antiviral, and antitumor effects (Fensterl and Sen, 2009).

Standard interferon (interferon α-2b) was the first drug to be approved for the treatment of hepatitis B. Long-term follow-up studies of interferon responders showed that it reduces the risk of progression to cirrhosis and HCC for both HBeAg-positive and HBeAg-negative carriers (Papatheodoridis et al., 2001; van Zonneveld et al., 2004). Nowadays pegylated interferon α-2a has superseded standard interferon due to higher response rates and the easier dosing schedule of pegylated interferon (weekly administration instead of daily or three times a week administration). The process of pegylation adds polyethylene glycol to interferon, which enhances the half-life compared to native interferon (Lau et al., 2005; Marcellin et al., 2004).

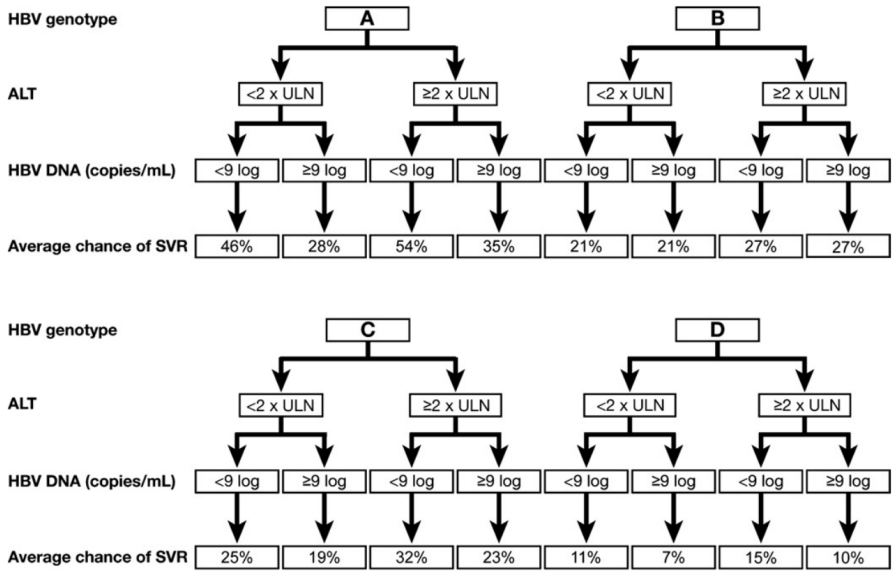

Figure 2.11: The diagram allows to compute the expected probability of sustained response to interferon for HBeAg-positive patients. Permission to use this figure was granted by Elsevier (Buster et al., 2009).

Standard therapy with pegylated interferon α-2a consists of one year of weekly administration. Treatment response is evaluated 24 weeks after the administration of the last dose, after the so-called follow-up period. Approximately 30% of HBeAg-positive patients achieve HBeAg seroconversion during the course of the treatment (Chan et al., 2005; Janssen et al., 2005; Lau et al., 2005). A long-term follow-up study of 172 patients with a mean observation period of three years revealed HBeAg and HBsAg seroconversion rates of 37% and 11%, respectively (Buster et al., 2008).

For HBeAg-negative patients the rates of sustained virological response (suppression of HBV DNA to below 400 copies per milliliter) were 19% after one year of treatment with pegylated interferon α-2a and 24 weeks of follow-up (Marcellin et al., 2004). After 3 years of long-term follow-up normal ALT levels were observed in 31% of the patients and 28% of the cohort showed low HBV DNA levels (below 10,000 copies per milliliter). Loss of HBsAg, thus cure of hepatitis B, was observed in 8.7% of the cohort.

Interferon facilitates limited treatment durations, but has partly severe side effects. Almost all patients experience initial flu-like illness, fatigue, and anorexia. Other common adverse effects include emotional unstableness, hair loss, and exacerbation of autoimmune illnesses (Kwon and Lok, 2011). Interferon treatment might accelerate liver inflammation, thus, it is contraindicated for patients with decompensated cirrhosis or HBV-related liver failure (Hoofnagle et al., 1993; Kwon and Lok, 2011). Nevertheless, interferon is effective and safe for patients with compensated cirrhosis.

Low response rates between 19% and 30% in combination with painful and severe side effects make the prediction of treatment response to interferon based on pretreatment patient or virus characteristics a clinically relevant research question. High pretreatment serum alanine aminotransferase levels, low HBV DNA levels, the presence of HBV genotypes A

and B, and progressed liver inflammation have been associated with higher response rates for interferon treatment (Lok and McMahon, 2001; Niederau et al., 1996). Buster et al. (2009) derived an algorithm to compute the expected probability of treatment response to pegylated interferon based on the treatment history of 721 HBeAg-positive patients (Figure 2.11). Nevertheless, pretreatment parameters provide only limited evidence of treatment response. Recent research led to the development of so-called response-guided interferon therapy. During treatment, decline of hepatitis B surface antigen and HBV DNA at week twelve were shown to be effective predictors of nonresponse to interferon (Chan et al., 2011; Rijckborst et al., 2010, 2012; Sonneveld et al., 2010, 2013). If no significant decline of quantitative HBsAg levels and HBV DNA is achieved during the first twelve weeks of treatment the chances of sustained response are near zero, thus treatment can be stopped. The underlying determinants of response to interferon, however, may be genetic factors and immunological disposition of the host as well as genetic factors of the virus itself beyond HBV genotype (Kamatani et al., 2009).

2.3.2 Treatment with Nucleos(t)ide Analogs

Nucleos(t)ide analogs are the second major treatment option for chronic hepatitis B. These compounds interfere with the HBV replication cycle by inhibiting the reverse transcription. Nucleos(t)ide analogs are chemically similar to deoxynucleotides, the natural building blocks of DNA. As a consequence the HBV polymerase incorporates nucleos(t)ide analogs into the DNA chain during reverse transcription of the pregenomic RNA into viral DNA. Nucleos(t)ide analogs lack a 3′-hydroxyl group, thus they do not allow the formation of a further 5′-3′ phosphodiester bond needed to extend the DNA chain (chain termination). This prevents the synthesis of the HBV DNA.

Treatment with nucleos(t)ide analogs effectively blocks viral replication, which lowers the long-term risk of liver failure. However, this effect lasts only as long as HBV has not developed resistance to the treatment. Resistance is mediated by suitable mutations in the reverse transcriptase domain of the polymerase gene. A resistant form of the polymerase can distinguish nucleos(t)ide analogs from natural deoxynucleotides and incorporates nucleos(t)ide analogs at lower rates compared to wild-type polymerases. Resistance is the major problem in the long-term treatment of hepatitis B with nucleos(t)ide analogs. Resistance becomes evident as virological breakthrough, which is defined as an increase of the HBV DNA level of at least one $\log_{10}$ level or the redetection of HBV DNA after it had become undetectable (Kwon and Lok, 2011). Increase of alanine aminotransferase levels (biochemical breakthrough) and increased liver inflammation may follow virological breakthrough if therapy is not adapted.

The five approved nucleos(t)ide analogs are classified into three groups based on their chemical structure.

- **L-nucleosides.** Lamivudine (3TC) and Telbivudine (LdT) are enantiomers of the natural nucleosides and have an inverted configuration at all chiral centers. It was demonstrated that viral polymerases process L-nucleosides better than the respective d-enantiomeric counterparts.

- **Acyclic nucleoside phosphonates.** Adefovir (ADV) and Tenofovir (TDF) have

Amino acid substitutions in the RT domain	3TC	LdT	ETV	ADV	TDF
M204I	R	R	I	S	S
L180M+M204V	R	R	I	S	S
N236T	S	S	S	R	I
A181T/V	I/R	R	S	R	I
L180M+M204I/V±I169T±V173L±M250V	R	R	R	S	S
L180M+M204I/V±T184G±S202I/G	R	R	R	S	S

Table 2.3: Drug resistance mutations of all approved nucleos(t)ide analogs with reference to Zoulim et al. (2009). R indicates resistance, I intermediate resistance, and S indicates susceptibility. Row five (L180M+M204I/V±I169T±V173L±M250V) and row six (L180M+M204I/V±T184G±S202I/G) list two disjunct ETV resistance pathways. At least one of the mutations I169T, V173L, M250V, T184G, or S202I/G together with L180M and M204I/V is required to confer full clinical resistance towards ETV. Positions are numbered with respect to the reverse transcriptase domain of the HBV polymerase according to the standard notation (Stuyver et al., 2001).

to be phosphorylated (intracellularily by cellular kinases) before they can interact with the HBV polymerase.

- **D-cyclopentane.** Entecavir (ETV) has a D-configured cyclopentyl group that fits directly into the hydrophobic pocket of HBV polymerase.

Cross resistance is defined as resistance to drugs to which a virus has never been exposed. Resistance to any drug confers at least some degree of cross resistance to other members of its group, and may reduce effectiveness of nucleos(t)ide analogs from other groups. Lamivudine and Telbivudine have very similar sets of resistance mutations (Table 2.3). A single nucleotide change is required to confer full clinical resistance. The two most important resistance mutations rtM204I/V (Isoleucine or Valine at position 204 of the reverse transcriptase) that confer resistance to Lamivudine and Telbivudine reduce sensitivity to Entecavir but not to Adefovir or Tenofovir. Multiple mutations in addition to rtM204I/V are required for high-level resistance to Entecavir (Table 2.3). The resistance mutations rt181T/V are shared between L-nucleosides and the acyclic nucleoside phosphonate Adefovir. The rtN236T mutation, which confers resistance to Adefovir and diminishes sensitivity towards Tenofovir, does not confer significant cross resistance to L-nucleosides and Entecavir. Cross resistance is held responsible for the high rate of drug resistance in patients with previous treatment failure (Rapti et al., 2007; Tenney et al., 2009). As a consequence, it is advised to start therapy with Tenofovir and Entecavir (European Association for the Study of the Liver, 2012). These drugs are very potent inhibitors of the viral replication and require multiple nucleotide exchanges to develop resistance. This reduces the long-term risk of drug resistance in hepatitis B patients.

NAs often have to be administered life-long (Kwon and Lok, 2011). Finite treatment duration might be achievable for HBeAg-positive patients who experience HBeAg seroconversion during NA administration. Treatment should be continued for at least 12 months after HBeAg seroconversion though. Rates of sustained response (defined as HBeAg-negativity

and undetectable HBV DNA levels) were found to be as high as 91% after five years of follow-up if NA treatment was continued for at least 12 months after HBeAg seroconversion (Lee et al., 2010). However, treatment duration with NAs is unpredictable prior to therapy as it depends on the timing of HBeAg seroconversion.

One year of NA treatment results in undetectable HBV DNA levels in up to 93% of HBeAg-negative patients (Lai et al., 2007; Marcellin et al., 2008b). Nevertheless, viral relapse occurs in almost all patients only 24 weeks after the discontinuation of NA administration (Lai et al., 2006). After four to five years of NA treatment 67% to 86% of HBeAg-negative patients show maintained viral suppression after discontinuation of treatment (Hadziyannis et al., 2006; Marcellin et al., 2010). Thus, a treatment pause with tight patient monitoring might be scheduled after four to five years of NA administration to evaluate off-treatment suppression of viral replication.

In contrast to treatment with interferon, treatment with nucleos(t)ide analogs is tolerated quite well (Kwon and Lok, 2011). Adefovir and Tenofovir are known to be associated with nephrotoxicity and have been reported to cause renal tubular dysfunction, including Fanconi syndrome (Heathcote et al., 2011; Marcellin et al., 2008a). Children treated with Tenofovir experienced decrease in bone mineral density (Purdy et al., 2008). One case of lethal lactic acidosis was reported in a patient who had highly impaired liver function before treatment (Lange et al., 2009). Nevertheless, nucleos(t)ide analogs show convincing short-term safety. The long-term safety in terms of decades of administration of these drugs has obviously not been established yet.

2.4 DNA Sequencing

DNA sequencing is the process of measuring the order of the four types of nucleotides along a DNA chain molecule. It is the key technology to determine the quasispecies of HBV. Sanger sequencing (first-generation sequencing) refers to a set of sequencing technologies, which utilize dideoxynucleotides (chain terminator nucleotides) in a polymerase chain reaction (PCR) to obtain DNA fragments of different lengths, which are separated by electrophoresis (Sanger et al., 1977a). Frederick Sanger, who was honored with the Nobel Prize in Chemistry in 1980 for his work on DNA sequencing, derived the first full genome DNA sequence of an organism called bacteriophage ϕX174 using this approach (Sanger et al., 1977b). Later in 2000, the first draft of the human genome was published by the use of Sanger sequencing technologies, which by the time had undergone several significant improvements (Lander et al., 2001; Venter et al., 2001). In the mid-2000's, second-generation sequencing (2ndGS) technologies became available and replaced Sanger sequencing in many applications (Mardis, 2008; Metzker, 2010). With respect to sequencing of viral populations, Sanger sequencing has several significant limitations: sensitivity to detect minor variants and the loss of linkage information. Second-generation sequencing technologies make use of high levels of miniaturization and parallelization to increase the sequencing throughput and to afford deeper insights into the composition of the viral quasispecies. Moreover, second-generation sequencing technologies amplify and read-out individual strands of DNA and, thus provide linkage information over the full read length. Nevertheless, as of 2014, Sanger sequencing is still used in the clinical routine for the treatment of hepatitis B due to its wide availability and low cost. In the remainder of

this Section we briefly introduce the cornerstones of modern Sanger sequencing and two second-generation sequencing technologies, Roche/454 and Illumina/MiSeq.

2.4.1 Sanger Sequencing

The following outline takes reference to Janitz (2011). Sanger sequencing utilizes a PCR to facilitate sequencing (Figure 2.12). A *primer* specific to the DNA substratum to be sequenced is required to initiate the PCR. The primer is an oligonucleotide complementary to a designated region of the DNA substratum. Double-stranded DNA needs to unwind and separate into single-stranded DNA. This allows the primer to bind to the designated region, which may be on either of the two complementary DNA strands. The separation of the two complementary DNA strands is referred to as DNA denaturation, which is achieved by heating (with a temperature high enough to break the hydrogen bonds between the complementary strands). In the subsequent PCR the primer is extended by incorporation of nucleotides according to the DNA substratum. The trick is that a small portion of dideoxynucleotides in addition to normal nucleotides is included in the reaction. Dideoxynucleotides are characterized by the absence of the 3′-hydroxyl group, which prevents the binding of another nucleotide at the 3′ end as no phosphodiester bond can be established (see upper-right corner of Figure 2.12). Dideoxynucleotides are incorporated during the PCR with a certain rate, which is mainly determined by the biochemical structure of the DNA polymerase that facilitates the PCR, the DNA sequence context, and the chemical structure of the dideoxynucleotides. The sporadic incorporation of dideoxynucleotides leads to chain termination at different sequence positions and results in DNA fragments of different lengths.

Dideoxynucleotides are either labeled with radioactive or fluorescent tags. The latter technique was introduced in 1986 and allows the PCR to be performed in a single reaction tube (Smith et al., 1986). Sequencing with radioactive labeled dideoxynucleotides requires four separate reactions, each enriched with one of the four dideoxynucleotides to determine the sequence information of the respective base. Radioactively tagged dideoxynucleotides are localized by X-rays, which stimulate the emission of beta particles. Likewise, a laser excites the fluorescently tagged dideoxynucleotides and the corresponding light signal is recorded by a camera module.

Before 1990, DNA fragments of different lengths were separated by electrophoresis using polyacrylamide gel (PAG) as a medium. The principle of electrophoresis says that charged particles migrate in an electric field with smaller fragments migrating faster. Thus, the negatively charged DNA fragments are separated by size. Nowadays, capillary electrophoresis affords faster separation of DNA fragments with higher resolution and better separation efficiency (Guttman et al., 1990; Luckey et al., 1990; Swerdlow et al., 1990). Capillary electrophoresis is a miniaturized version of electrophoresis using PAG in a capillary, which minimizes disturbance by optimizing the surface-area-to-volume ratio.

Another cornerstone of efficient Sanger sequencing is cycle sequencing introduced by Murray (1989). In cycle sequencing the first steps of the sequencing process (denaturation, primer binding, and chain extension/termination) are repeated 20 to 50 times by cycling through distinct temperatures optimized to support the respective reactions of the sequencing process. This improves the sensitivity of the sequencing reaction and permits

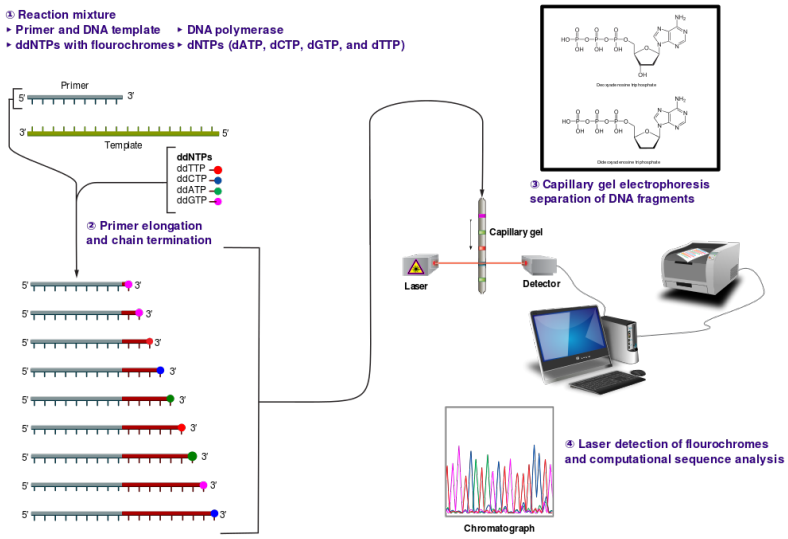

Figure 2.12: Visualization of the four steps of Sanger sequencing. (1) The sequencing reaction mixture includes the DNA substratum to be sequenced, primer to initiate the polymerase chain reaction, normal nucleotides (dNTPs), and dideoxynucleotides (ddNTPs). (2) Sporadic incorporation of dideoxynucleotides during PCR leads to chain termination at different positions and DNA fragments of variable length. (3) Variable-length DNA fragments are separated by size by the use of capillary gel electrophoresis. (4) Fluorescently tagged dideoxynucleotides are detected by laser/camera aperture. Fluoresce data is read-out as a chromatogram. Subsequent computational analysis of the chromatogram provides the DNA sequence. Figure was obtained from Wikipedia (`http://en.wikipedia.org/wiki/File:Sanger-sequencing.svg`).

sequencing of small amounts of DNA substrata.

Other very important lines of research that have led to the excellent effectivity and efficiency of Sanger sequencing today is based on the development of optimized fluorescent dyes (Ju et al., 1995; Metzker et al., 1996; Rosenblum et al., 1997), optimized dideoxynucleotides (Prober et al., 1987), and polymerases specifically designed for sequencing (Reeve and Fuller, 1995; Tabor and Richardson, 1989, 1990). All components of the chemistry are optimized among other properties to provide high physio-chemical stability, equal incorporation rates of the four dideoxynucleotides, minimally overlapping emission spectra, high fluorescence intensity, and uniform electrophoretic mobilities. The importance of the optimization of the chemistry to allow efficient read-out of the fluorescence intensity data and its subsequent transformation into sequence data can not be overestimated.

The raw image data need to be preprocessed to derive sequencing chromatograms, which consist of deconvolved, smoothed, and noise-reduced fluorescence intensity values at several thousand uniformly spaced time points. The sequencing chromatograms are then translated into a DNA sequence, referred to as base-calling. Base-calling is algorithmically trivial for ideal data, which is noise-free and in which all peaks are evenly spaced and non-overlapping. But several experimental and systematic factors lead to impaired and noisy chromatograms, which are more difficult to interpret (see Ewing et al. (1998) for a summary of influencing factors). Progress was made in the development of efficient and error-tolerant algorithms and their implementations to derive and interpret sequencing chromatograms (Ewing et al., 1998; Flood et al., 2002; Nickerson et al., 1997).

2.4.2 Roche/454 Pyrosequencing

The Roche/454 FLX sequencer was the first commercially available second-generation sequencing technology (Margulies et al., 2005). It performs several hundred thousand independent pyrosequencing reactions in the wells of a picotiter plate. We now briefly introduce pyrosequencing, a sequencing technology that was developed in the 1990s, and the Roche/454 miniaturization technique with reference to Mardis (2008).

Pyrosequencing. Pyrosequencing is an implementation of the idea of sequencing-by-synthesis rather than sequencing by chain termination (Ronaghi et al., 1996, 1998). Given a single-stranded DNA stratum, the complementary strand is synthesized by repetitively adding only one of the four possible natural nucleotides at a time. Incorporation of nucleotides by a DNA polymerase results in the release of pyrophosphate, which is converted to adenosine triphosphate (ATP) by an ATP sulfurylase. ATP is made observable as light using the firefly luciferase. The number of nucleotides that were added in a row determines the amount of produced ATP and, thus the intensity of the light emitted. The interpretation of the sequence of light signals of varying intensity affords the read-out of the DNA sequence.

Roche/454 pyrosequencing. The pyrosequencing technology was refined and improved by Roche/454 to allow massively parallel sequencing. In the library preparation preprocessing step, the DNA strata to be sequenced are marked with individual 454-specific adapter sequences. The sequencing strata are then mixed with agarose beads, which carry oligonucleotides complementary to the 454-specific adapter sequences on their surfaces. Each bead is associated with at most one DNA strand of the sequencing stratum. The re-

sulting complexes of beads and DNA are then isolated into individual water-in-oil emulsion micelles. Inside the micelles a PCR is initiated that results in approximately one million copies of each individual DNA strand. The beads are then single-fitted into the wells of a picotiter plate, where individual pyrosequencing reactions are initiated, maintained, and monitored. The wells ensure that the beads are fixed to a specific location on the plate. This allows for the recording of the individual fluorescent emissions of the individual pyrosequencing reactions and facilitates the determination of the DNA sequences.

2.4.3 Illumina/MiSeq Sequencing

Illumina/MiSeq utilizes a sequencing-by-synthesis approach that makes use of custom chain terminator nucleotides, derivatives of natural nucleotides in which the 3′-hydroxyl group is chemically blocked. The blocking of the 3′-hydroxyl group ensures that only one nucleotide is added at a time. The chain terminator nucleotides are fluorescently labeled such that the type of nucleotide can be identified upon incorporation. Sequencing takes place in turns. Each turn consists of three steps. First, all four types of chain terminator nucleotides are added. Second, incorporation is recorded in the imaging step. Third, the 3′ blocking component is chemically removed such that further chain extension is feasible.

Before the sequencing reaction, the DNA strata need to be amplified in the so-called cluster creation phase. Therefore, single-stranded DNA fragments, which were randomly fragmented and ligated with adapter sequences at both ends, are immobilized on the surface of a flow cell. Then, solid-phase amplification is initialized that creates approximately one million copies of each DNA fragment in close proximity to each other (clusters of identical DNA fragments). Solid-phase amplification was developed as a novel method to amplify DNA, in which primers are attached to a solid surface (Adessi et al., 2000; Bing et al., 1996).

2.5 Probabilistic Reasoning and Statistical Learning

Throughout the thesis we make use of probabilistic reasoning and statistical learning methods to infer information from data. In the following Sections we briefly introduce the basic rules and notations of probability theory, probabilistic reasoning, and two important statistical learning models.

2.5.1 Probability Theory and Probabilistic Reasoning

The following introduction to probability theory and probabilistic reasoning has been adapted from Barber (2012).

Random variables are used to describe possible outcomes of random experiments, e.g. the experiment of a coin toss, which can result in either “heads” or “tails”. The set of all possible outcomes or states of a random variable x is referred to as its domain: $\mathrm{dom}(x)$. For now we assume $\mathrm{dom}(x)$ to be a finite set. The probability of a random variable x to be in state $a \in \mathrm{dom}(x)$ is denoted by $P(x = a)$. $P(x = a)$ takes values between 0 and 1 and quantifies the level of belief that x is in state a. $P(x = a) = 1$ means that we are absolutely certain that x is in state a and $P(x = a) = 0$ means that we are absolutely

certain that x is not in state a. For $P(x = a)$ we use the short notation $P(x)$ to depict the probability of any respective state $a \in \mathrm{dom}(x)$.

The normalization condition ascertains that the probabilities over all states add up to one:

$$\sum_{a \in \mathrm{dom}(x)} P(x = a) = 1.$$

To sum over all possible states of a random variable we usually write $\sum_x P(x)$ instead of $\sum_{a \in \mathrm{dom}(x)} P(x = a)$.

Additionally, the rules of probability theory states that for two random variables x and y the following equation holds:

$$P(x \text{ and } y) = P(x) + P(y) - P(x \text{ or } y).$$

The notation $P(x, y)$ is used as shorthand to depict $P(x \text{ and } y)$.

Based on the *joint distribution* $P(x_1, \ldots, x_n)$ of a set of random variables the distribution of a subset of these variables can be computed by the process of *marginalization*.

$$P(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = \sum_{x_i} P(x_1, \ldots, x_n) \tag{2.1}$$

After $n-1$ steps of marginalization in which all variables except x_i have been marginalized $P(x_i)$ denotes the *marginal distribution* of x_i with respect to the joint distribution $P(x_1, \ldots, x_n)$.

If $dom(x)$ is finite, x is called a discrete random variable. For a real-valued random variable x ($dom(x) = \mathbb{R}$) a *probability density function* is defined as a function

$$f(x) : dom(x) \longmapsto \mathbb{R}$$

such that

$$f(x) \geq 0,$$
$$\int_{-\infty}^{\infty} f(x)dx = 1,$$

and the probability that x takes a value between $a \in \mathbb{R}$ and $b \in \mathbb{R}$ is given by

$$P(a \leq x \leq b) = \int_a^b f(x)\ dx.$$

The multivariate case $f(x_1, \ldots, x_n)$ is treated analogously with integration over the respective regions to compute $P(a_1 \leq x_1 \leq b_1, \ldots, a_n \leq x_n \leq b_n)$. Note that formally $P(x = a)$ is zero for any real-valued variable x and all $a \in \mathbb{R}$. However, we write $P(x)$ for both discrete and real-valued variables, thus we do not distinguish between probabilities and probability density function values. In the real-valued case the expression $P(x)$ can be considered as a shorthand for $\int_{x \in \Delta} f(x)\ dx$ where Δ is a small region centered around x. This is well defined in a probabilistic sense and in the limit (Δ being very small), this would give $P(x) \approx \Delta f(x)$. The same Δ is then consistently used for all probability density functions, which results in a common prefactor Δ in all expressions. Usually the Δ values can be ignored as they cancel out when computing relative probabilities. In this way, the standard rules of probability carry over to real-valued variables. Further, we write

$\int_x P(x)\ dx$ as a valid expression for real-valued and discrete random variables x assuming that integrations are replaced by finite summations for discrete random variables.

Next, we introduce *conditional probabilities* $P(x|y)$. $P(x|y)$ denotes the probability of x given that we know the state of y. If $P(y) > 0$ the conditional probability $P(x|y)$ is defined as:

$$P(x|y) = \frac{P(x,y)}{P(y)}. \tag{2.2}$$

From this definition and the fact that $P(x,y) = P(y,x)$ Bayes' rule is derived:

$$P(x|y) = \frac{P(x)P(y|x)}{P(y)}.$$

Bayes' rule. Bayes' rule plays a key role in probabilistic reasoning. It allows us to relate $P(x|y)$ with $P(y|x)$. This is useful in the context of a *generative model* $P(\mathcal{D}|\theta)$ of the data $\mathcal{D}$, which depends on a set of model parameters θ. The generative model $P(\mathcal{D}|\theta)$ is a hypothesis about how the data $\mathcal{D}$ was sampled dependent on θ. Using Bayes' rule we compute $P(\theta|\mathcal{D})$, which summarizes our knowledge about the distribution of θ after we have observed data $\mathcal{D}$:

$$P(\theta|\mathcal{D}) = \frac{P(\theta)P(\mathcal{D}|\theta)}{P(\mathcal{D})}.$$

In this context we call $P(\theta|\mathcal{D})$ the *posterior*, $P(\theta)$ the *prior*, $P(\mathcal{D}|\theta)$ the *data likelihood*, and $P(\mathcal{D})$ the *evidence*. The evidence $P(\mathcal{D})$ is defined such that $P(\theta|\mathcal{D})$ is a valid probability distribution and sums up to 1. Thus,

$$P(\mathcal{D}) = \int_\theta P(\theta)P(\mathcal{D}|\theta)\ d\theta.$$

The process of probabilistic reasoning may be reduced to the computation of θ^*, which is the value of θ that maximizes the posterior distribution:

$$\theta^* = \arg\max_\theta P(\theta|\mathcal{D}).$$

θ^* is the so-called *maximum a posteriori probability* (MAP) estimate, which is an reasonable estimate for θ incorporating prior knowledge and observed data.

The application of Bayes' rule requires the quantification of the prior belief of the distribution $P(\theta)$. If $P(\theta)$ assigns equal probability to all values of θ (a so-called flat prior) the MAP estimate reduces to the *maximum likelihood* estimate, which solely maximizes the data likelihood $P(\mathcal{D}|\theta)$.

Model selection. Bayes' rule is also applied when comparing alternative generative models $P(\mathcal{D}|M_1,\theta_1), \ldots, P(\mathcal{D}|M_m,\theta_m)$. Each model M_i has a prior probability $P(M_i)$ and its own set of parameters θ_i with a prior distribution $P(\theta_i|M_i)$. We are interested in the models' posterior probabilities, which are given by

$$P(M_i|\mathcal{D}) = \frac{P(M_i)P(\mathcal{D}|M_i)}{\sum_{j=1}^m P(M_j)P(\mathcal{D}|M_j)}$$

where

$$P(\mathcal{D}|M_i) = \int_{\theta_i} P(\theta_i|M_i)P(\mathcal{D}|M_i,\theta_i)\, d\theta_i. \tag{2.3}$$

Two models M_i and M_j are compared by computing the *posterior Bayes' factor* which is the ratio of the models' posterior probabilities:

$$\frac{P(M_i|\mathcal{D})}{P(M_j|\mathcal{D})} = \frac{P(M_i)}{P(M_j)} \cdot \frac{P(\mathcal{D}|M_i)}{P(\mathcal{D}|M_j)}.$$

The quantities $P(\mathcal{D}|M_i)$ are called the marginal model likelihoods. In case of flat model priors $P(M_i)$, model selection is reduced to comparing the marginal model likelihoods, which do account for model complexity by averaging $P(\theta_i|M_i) \cdot P(\mathcal{D}|M_i,\theta_i)$ over the whole parameter space of θ_i (equation (2.3)). Model selection in Chapter 4 is based on the marginal model likelihoods.

To summarize, probabilistic reasoning requires a two step approach. First, one considers a finite set of models $M_1,\ldots,M_m$. Each model is specified by the model prior $P(M_i)$, the data likelihood function $P(\mathcal{D}|M_i,\theta_i)$, and a prior for the model parameters $P(\theta_i|M_i)$. Model selection is carried out by computing the models' posterior probabilities $P(M_i|\mathcal{D})$. These quantities are often computationally expensive to obtain as their computation requires to integrate over the set of model parameters according to equation (2.3). The model M_i that maximizes $P(M_i|\mathcal{D})$ might be selected, but $P(M_i|\mathcal{D})$ is not straightforward to interpret. Model posterior probabilities have to be interpreted in the context of the finite set of models $M_1,\ldots,M_m$ that were considered from the beginning. Ideally, one would need to consider all possible models and compute the models' posterior probabilities based on this infinite set of models. This, of course, is usually not feasible. Therefore, certainty of model selection is expressed in terms of posterior Bayes' factors, which are easier to interpret. In the second step of probabilistic reasoning, Bayes' rule allows us to infer the posterior distribution of the respective model parameters.

2.5.2 Hidden Markov Models

Hidden Markov models (HMMs) are specific forms of probabilistic models that have been used extensively in bioinformatics. HMMs can be applied to infer the sequence of internal (hidden) states using a sequence of observed symbols. In particular, they have been used in protein secondary structure prediction (Won et al., 2007), gene prediction (Munch and Krogh, 2006), pairwise and multiple sequence alignments (Durbin et al., 1998; Pachter et al., 2002), and many more applications (Yoon, 2009). We employ HMMs in Chapter 3 to infer the HBV genotypes and the recombination breakpoints of HBV sequence data. In the following, we briefly lay out the formal representation of HMMs and the algorithms required to perform inference in HMMs. We stick close to the systematic and the notation used in Durbin et al. (1998).

Suppose we are given a finite sequence of hidden states $\pi = \pi_1,\ldots,\pi_L$ with $\pi_i \in \mathfrak{A}$ and a finite sequence of observed symbols $x = x_1,\ldots,x_L$ with $x_i \in \mathfrak{B}$, where each state or symbol belongs to a respective finite set $\mathfrak{A}$ or $\mathfrak{B}$. The sequence π is assumed to have the Markov property: the probability of π_i only depends on the previous state π_{i-1}. The transition from π_{i-1} to π_i is characterized by a set of parameters a_{kl}.

$$P(\pi_i = l|\pi_{i-1} = k) = a_{kl}$$

Additional parameters are the transition probabilities a_{0k} from a virtual start state 0 to state k. In other words, a_{0k} is the probability to start in state k.

It is further assumed that the observed symbol x_i only depends on the hidden state π_i via so-called emission probabilities. The emission probabilities $e_k(b)$ are additional parameters of the model that specify the conditional probabilities $P(x_i|\pi_i)$.

$$P(x_i = b|\pi_i = k) = e_k(b)$$

Based on this model description, several distinct statistical inference problems can be addressed. In the remainder of this Section, we lay out the statistical problems together with the respective algorithms to perform efficient inference.

input : $e_k(b), a_{kl}, x_i$
output: π^*
Initialization:
$v_k(0) = a_{0k}$ for $k \in \mathfrak{A}$
Recursion:
for $i = 1, \ldots, L$ **do**
  $v_l(i) = e_l(x_i) \max_k(v_k(i-1)a_{kl})$, for all $l \in \mathfrak{A}$
  $\text{ptr}_i(l) = \arg\max_k(v_k(i-1)a_{kl})$, for all $l \in \mathfrak{A}$
end
Termination:
$\pi^*_L = \arg\max_k(v_k(L))$
Traceback:
for $i = L, \ldots, 1$ **do**
  $\pi^*_{i-1} = \text{ptr}_i(\pi^*_i)$
end

Algorithm 1: Viterbi algorithm

Optimal hidden state sequence. A very important statistical inference problem is to find the most probable sequence of hidden states π^* given a sequence of symbols x, the transition probabilities a_{lk}, and the emission probabilities $e_k(b)$. More formally, we are interested in:

$$\pi^* = \arg\max_{\pi} P(\pi|x).$$

π^* can be computed using the Viterbi algorithm, a dynamic programming algorithm. The key idea is very simple. Let $v_l(i)$ be the probability of the most probable subsequence $\pi_1, \ldots, \pi_i$ such that $\pi_i = l$. Then, for any state l the quantities $v_l(i+1)$ can be computed using all predecessor states $k \in \mathfrak{A}$:

$$v_l(i+1) = e_l(x_{i+1}) \max_k v_k(i)a_{kl}. \tag{2.4}$$

Starting with $v_l(0) = a_{0k}$, the most probable path $v_l(L)$ ending at the last position L in state l can be computed with L iterations using equation 2.4. Then, π^*_L is given by

$\arg\max_k v_k(L)$ and π^*_{i-1} for $i = L, \ldots, 2$ is given by the backtracking pointers $\text{ptr}_i(\pi^*_i)$, which are computed simultaneously with $v_l(i)$. The Viterbi algorithm is listed as Algorithm 1.

Sequence probability. To compute the probability $P(x)$ of a sequence of symbols x, the different paths of states π need to be taken into account. According to the marginalization rule of probability (equation 2.1):

$$P(x) = \sum_{\pi} P(x, \pi).$$

The size of the summation grows exponentially with the length of the sequence x. Thus, enumeration of all possible paths π does not allow for efficient computation of $P(x)$. The forward algorithm can compute $P(x)$ in polynomial time. It is a slight modification of the Viterbi algorithm, in which the maximization steps are replaced with summation steps. Let $f_l(i)$ be the probability of the subsequence $x_1, \ldots, x_i$ with $\pi_i = l$.

$$f_l(i) = P(x_1, \ldots, x_i, \pi_i = l)$$

Then, $f_l(i+1)$ can be computed by summation over the quantities $e_l(x_{i+1}) f_k(i) a_{kl}$ for all predecessor states $k \in \mathfrak{A}$. The forward algorithm is detailed as Algorithm 2.

input : $e_k(b), a_{kl}, x_i$
output: $P(x), f_k(i)$
Initialization:
$f_k(0) = a_{0k}$ for $k \in \mathfrak{A}$
Recursion:
for $i = 1, \ldots, L$ **do**
  $f_l(i) = e_l(x_i) \sum_{k \in \mathfrak{A}} f_k(i-1) a_{kl}$, for all $l \in \mathfrak{A}$
end
Termination:
$P(x) = \sum_{k \in \mathfrak{A}} f_k(L)$

Algorithm 2: Forward algorithm

Posterior state probability. The computation of the most probable path π^* using the Viterbi algorithm does not provide uncertainty information. The position-specific posterior probability of a state l given the observed sequence $P(\pi_i = l|x)$ quantifies the belief that observation x_i originating from state k and is of interest in many applications. The position-specific posterior can be computed by the forward-backward algorithm. This algorithm requires the quantities $f_l(i)$ derived from the forward algorithm and the quantities $b_l(i)$ computed by the backward algorithm. The term $b_l(i)$ is defined as the probability of the subsequence $x_{i+1}, \ldots, x_L$ given that x_i is emitted from state l:

$$b_l(i) = P(x_{i+1}, \ldots, x_L | \pi_i = l).$$

The quantities $b_l(i)$ are computed backwards (from $i = L-1$ to $i = 1$) using a summation recursion formula (see Algorithm 3).

input : $e_k(b), a_{kl}, x_i$
output: $P(x), b_l i$
Initialization:
$b_k(L) = 1$ for $k \in \mathfrak{A}$
Recursion:
for $i = L-1, \ldots, 1$ **do**
  $b_l(i) = \sum_{k \in \mathfrak{A}} a_{lk} e_k(x_{i+1}) b_k(i+1)$, for all $l \in \mathfrak{A}$
end
Termination:
$P(x) = \sum_{k \in \mathfrak{A}} a_{0k} e_k(x_1) b_k(1)$

Algorithm 3: Backward algorithm

Using the definition of conditional probabilities (equation 2.2) the posterior probability $P(\pi_i = l|x)$ can be expressed as

$$P(\pi_i = l|x) = \frac{P(x, \pi_i = l)}{P(x)}. \tag{2.5}$$

We rewrite $P(x, \pi_i = l)$ using, first, the definition of conditional probabilities and, second, the Markov property that the probability of the subsequence $x_{i+1}, \ldots, x_L$ depends only on the state π_i. Thus,

$$\begin{aligned} P(x, \pi_i = l) &= P(x_1, \ldots x_i, \pi_i = l) P(x_{i+1}, \ldots, x_L | x_1, \ldots x_i, \pi_i = l) \\ &= P(x_1, \ldots x_i, \pi_i = l) P(x_{i+1}, \ldots, x_L | \pi_i = l). \end{aligned} \tag{2.6}$$

Note that $P(x, \pi_i = l)$ is the product of $f_l(i)$ and $b_l(i)$. Together with equation (2.5) this gives rise to the forward-backward algorithm to compute the position-specific posterior.

$$P(\pi_i = l|x) = \frac{f_l(i) b_l(i)}{P(x)} \tag{2.7}$$

To sum up, the forward-backward algorithm requires the execution of the forward algorithm, the backward algorithm and, finally, the application of equation (2.7).

2.5.3 Support Vector Machines

Support vector machines (SVMs) are a set of supervised learning algorithms, algorithms to solve supervised learning problems. SVMs are very popular in bioinformatics and in many areas of machine learning due to their high prediction performance (Wang, 2005b). We use SVMs to predict treatment response to interferon in Chapter 6. The following outline is adapted from Hastie et al. (2009).

In supervised learning problems we are given joint observations of two random variables x and y, thus data points that consist of pairs $(x_1, y_1), \ldots, (x_n, y_n)$ with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. The general goal of supervised learning is to derive a function $f(x)$ that predicts the value of y given the value of x. We refer to $x_1, \ldots, x_n$ as the input (input data) of the learning algorithm and to $y_1, \ldots, y_n$ as the output (output data) or the labels. The p dimensions of $x_1, \ldots, x_n \in \mathbb{R}^p$ are called the features or the feature set.

Supervised learning problems are divided into two groups:

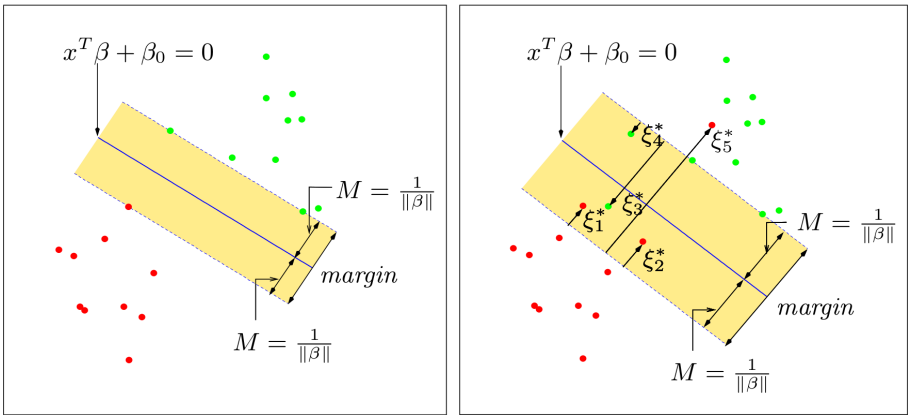

Figure 2.13: Support vector machines with perfect separation (left) and soft margin (right). Permission to reuse this figure was granted by Springer (Hastie et al., 2009).

(i) **Regression problems**, in which case the labels y are quantitative (continuous).

(ii) **Classification problems**, in which case the labels y are qualitative (categorical). The distinct values of y are called classes. Binary (or 2-fold) classification problems, in which y can take exactly two different values -1 and $+1$, are of special importance.

In the following we briefly describe linear soft-margin SVMs used for binary classification. The goal is to find a linear function

$$f(x) = \beta^T x + \beta_0$$

to estimate the label y. y is set to class -1 if $f(x) < 0$ and to class $+1$ otherwise. Suppose the input data of the two classes can be perfectly separated by many different hyperplanes (see left side of Figure 2.13). In such cases Vapnik (1996) suggested, in order to construct a more accurate classifier, to select the hyperplane which is maximally far away from any data point, the hyperplane which maximizes the so-called margin[7]. The concept of soft-margin SVMs generalizes this idea to the case of non-separable classes (right side of Figure 2.13). Here data points are allowed to reside on the wrong side of the hyperplane (Boser et al., 1992; Cortes and Vapnik, 1995). In order to find the optimal soft-margin separating hyperplane the following optimization problem needs to be solved.

$$\begin{aligned} \underset{\{\beta,\beta_0\}}{\text{minimize}} \quad & \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i(x^T\beta + \beta_0) \geq 1 - \xi_i, \quad i = 1,\ldots,n \\ & \xi_i \geq 0, \qquad\qquad\qquad\quad i = 1,\ldots,n \end{aligned} \tag{2.8}$$

$C > 0$ denotes a regularization parameter which trades off model complexity for prediction accuracy on the training data. A solution is obtained by the use of the Lagrange multiplier method. This method introduces Lagrange multipliers $\alpha_i \geq 0$, $i = 1,\ldots,n$

[7]The margin is defined as the distance of the hyperplane to the closest point from either class.

and $\mu_i \geq 0,\ i = 1, \ldots, n$ associated each with one constraint, such that the problem reformulates to minimize the Lagrange primal function.

$$\begin{aligned} \underset{\{\beta,\beta_0,\alpha\}}{\text{minimize}} \quad & \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(x_i^T\beta + \beta_0) - (1-\xi_i)] - \sum_{i=1}^{n}\mu_i\xi_i \\ \text{subject to} \quad & \alpha_i \geq 0,\ i = 1, \ldots, n \\ & \mu_i \geq 0,\ i = 1, \ldots, n \end{aligned} \tag{2.9}$$

Setting the partial derivatives for β, β_0, and α to zero provides.

$$\beta = \sum_{i=1}^{n}\alpha_i y_i x_i \tag{2.10}$$

$$0 = \sum_{i=1}^{n}\alpha_i y_i \tag{2.11}$$

$$\alpha_i = C - \mu_i,\ i = 1, \ldots n \tag{2.12}$$

Equations (2.10) to (2.12) together with the positivity constraints $\alpha_i,\ \mu_i,\ \xi_i \geq 0$ can be substituted into (2.9) to obtain the Lagrange dual problem.

$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j \\ \text{subject to} \quad & \sum_{i=1}^{n}\alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C,\ i = 1, \ldots, n \end{aligned} \tag{2.13}$$

The Karush-Kuhn-Tucker conditions based on the Lagrange dual formulation provides the following constrains (Karush, 1939; Kuhn and Tucker, 1951).

$$0 = \alpha_i[y_i(x^T\beta + \beta_0) - (1-\xi_i)], \qquad i = 1, \ldots n \tag{2.14}$$

$$0 = \mu_i\xi_i, \qquad i = 1, \ldots n \tag{2.15}$$

$$0 \leq y_i(x^T\beta + \beta_0) - (1-\xi_i), \qquad i = 1, \ldots n \tag{2.16}$$

Finally, equations (2.10) to (2.12) together with equations (2.14) to (2.16) uniquely characterize the solution to the primal and dual problem and facilitate its efficient computation.

3 The Dual Infection Model

Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.

(George E.P. Box, 1987)

The impact of the HBV genotype on the natural progression of chronic hepatitis B and on treatment response is discussed in Sections 2.2.3 and 2.3. Whether or not to determine the genotype before start of therapy is a matter of debate (reviewed in Cooksley (2010); Lin and Kao (2011)). The three major regional liver associations, the American Association for the Study of Liver Disease (AASLD) (Lok and McMahon, 2009), the European Association for the Study of Liver (EASL) (European Association for the Study of the Liver, 2012), and the Asian Pacific Association for the Study of Liver (APASL) (Liaw et al., 2008) do not recommend genotyping in their respective guidelines. Nevertheless, several European treatment guidelines recommend genotyping before therapy, e.g. the German guidelines (Cornberg et al., 2011), the Swedish guidelines (Lindh et al., 2008), and, if treatment with interferon is considered, also the Dutch guidelines (Buster et al., 2012). Interferon treatment affords finite duration therapy of only 48 weeks, which is an important advantage compared to treatment with nucleos(t)ide analogues (Section 2.3). Sustained response to interferon strongly depends on the HBV genotype with which the patient is infected. Patients infected with genotype A and genotype B, the two genotypes that show good response to interferon, may benefit greatly from interferon therapy.

Several methods have been developed to determine the HBV genotype based on a patient serum sample. Phylogenetic analysis of the entire viral genome derived by population-based Sanger sequencing remains the gold standard. Additionally, other computational approaches were proposed to identify the genotype or a recombination of genotypes using full- or sub-genome sequence data (Alcantara et al., 2009; Myers et al., 2006; Rozanov et al., 2004; Schultz et al., 2006; Struck et al., 2010).

The importance and high prevalence of recombinant forms of the HBV genome is discussed in Section 2.2.3. Genotyping of recombinant forms have been addressed either by applying an existing genotyping method to individual sliding windows along the genome (Myers et al., 2006; Rozanov et al., 2004; Alcantara et al., 2009) or, more involved, by hidden Markov models (Schultz et al., 2006; Zhang et al., 2006). A prerequisite for recombination events is the simultaneous infection with two heterologous HBV strains referred to as HBV dual infection. Dual infections can be either coinfections or superinfections. Coinfection indicates a situation in which a patient was infected with two heterologous strains either simultaneously or within a brief period of time. According to the definition, the infection with the second strain occurs prior to the immune response to the first infection. Superinfection describes an infection with a second strain, which occurs after the initial infection and after immune response to it (Smith et al., 2005). Despite several studies on the natural progression of dual infections, their clinical implications are largely unknown

(Kao et al., 2001; Michitaka et al., 2005). Nevertheless, the importance of identifying and genotyping dual infections correctly to quantify the risk of interferon therapy failure was shown by Hannoun et al. (2002).

In this Chapter we focus on the identification and genotyping of intra- and intergenotype dual infections using population-based sequence data. In the presence of multiple heterologous viral strains population-based sequence data contains a high number of ambiguous sequence positions, which hamper state-of-the-art genotyping methods. The correct classification of patient sequence data is impaired and respective patients are at risk of receiving suboptimal treatments. We describe an extended version of the work presented in Beggel et al. (2012). The study was executed in cooperation with Jens Verheyen from the University of Cologne who provided the patient data from routine diagnostics and with Maria Neumann-Fraune from the University of Cologne who produced clonal sequence data to validate model predictions. This Chapter is structured as follows. Section 3.1 discusses *in silico* genotyping approaches and Section 3.2 summarizes *in vitro* methods to genotype HBV dual infections. Section 3.3 outlines the dual infection model and the validation procedure. Results on synthetic and patient data are presented in Sections 3.4 and 3.5, respectively. Section 3.6 provides a discussion and final remarks.

3.1 Genotyping HBV *in silico*

HBV genotypes are defined based on clusters in phylogenetic analysis of full genome sequences. However, the proper construction of phylogenetic trees is non-trivial, computationally intensive, and often requires visual inspection and verification. Thus, in application scenarios where thousands of sequences need to be genotyped, phylogenetic methods are not easy to be applied. Methods based on scoring functions were developed to overcome the limitations of phylogenetic methods and are described in the remainder of this Section. We discuss genotyping methods based on sequence similarity and those that utilize position-specific scoring matrices.

3.1.1 Genotyping by Sequence Similarity

A set of simple yet effective genotyping methods uses sequence similarity measures. Given a set of reference sequences for each genotype, the approach consists of computing the similarity of the input sequence to each reference sequence. The input sequence is then assigned to the genotype of the reference with highest sequence similarity measure unless the similarity score is less than a given threshold. Similarity measures often require a precomputed sequence alignment. The National Center for Biotechnology Information (NCBI) genotyping service bypasses this computationally expensive precondition by the utilization of BLAST similarity scores (Altschul et al., 1990; Rozanov et al., 2004). In order to determine and genotype recombinant forms, the sequence similarity methodology is applied to sliding windows along the input sequence. Rozanov et al. (2004) uses 300 bases long windows with an overlap of 100 bases. This gives rise to several windows covering each sequence position. Multiple genotype predictions may be processed to compute position-wise confidence scores.

3.1.2 Genotyping by Position-specific Scoring Matrices

Sequence similarity methods utilize only a small set of reference sequences. Detailed genetic information, for instance on conserved or variable regions in the viral genome, which nowadays can be extracted from large sequence databases, is neglected. A different set of approaches is based on position-specific scoring matrices. Large sets of reference sequences are employed to compute the distribution of nucleotides or amino-acids per sequence position with respect to a precomputed alignment. HBV STAR is based on position-specific distributions on the amino-acid level (Gale et al., 2004; Myers et al., 2005, 2006). HBV STAR genotypes an input sequence by rating each amino-acid position independently and summing up position-specific scores. Each amino-acid a_j in the amino-acid input sequence $a_i, \ldots, a_{i+n}$ is classified to either have a positive discriminant with respect to each genotype, if the odds ratio

$$\frac{\text{genotype-specific frequency of } a_j}{\text{genotype-independent frequency of } a_j \text{ (whole data set)}}$$

is greater than one or to have a negative discriminant otherwise. The quotient of the number of positive discriminants divided by the number of negative discriminants yields an overall discriminant odds ratio score for each genotype that summarizes all amino-acid positions. The genotype with the largest overall discriminant odds ratio score indicates the most likely genotype of the input sequence.

3.1.3 Detection of Recombinants by Position-specific Nucleotide Distributions

Position- and genotype-specific nucleotide distributions (PGSNDs) were combined in a concise probabilistic framework to genotype HBV, hepatitis C virus (HCV), and HIV (Schultz et al., 2006, 2009, 2012; Zhang et al., 2006). The basic idea reads as follows. Let $\{g_1, \ldots, g_l\}$ be a finite set of genotypes and let $\mathcal{A}$ be a multiple sequence alignment of a set of genotype-annotated reference sequences. The PGSNDs p_{gj}, where $g \in \{g_1, \ldots, g_l\}$ denotes the genotype and j denotes the sequence position within the multiple sequence alignment $\mathcal{A}$, are derived by counting base frequencies. The PGSNDs encode the distribution of genotype-specific polymorphisms at the nucleotide level and, thus, harbor information on conserved or variable regions within the genome. In this approach p_{gj} represent multinomial distributions over the alphabet $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ consisting of the four bases found in DNA Adenine, Cytosine, Guanine, and Thymine. The model assumes that $p_{gj}[b]$ is the probability of observing base $b \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ at sequence position j of an input sequence $r_i, \ldots, r_{i+n}$ (for which we assume it is aligned with respect to the alignment $\mathcal{A}$) subject to the condition that the sequence is of genotype g. Thus,

$$P(r_j = b|g) = p_{gj}[b].$$

Recombinant detection is implemented by modeling the genotype as the hidden state of a hidden Markov model. Figure 3.1 illustrate the so-called jumping profile hidden Markov model. The emission probabilities of the HMM are given by the PGSNDs p_{gj}. State/genotype transitions are allowed at any position in the model. Transitions between genotypes are referred to as jumps and allow for transitions if the input sequence shows local similarity to different genotypes. Jumps need to have relatively high transition costs

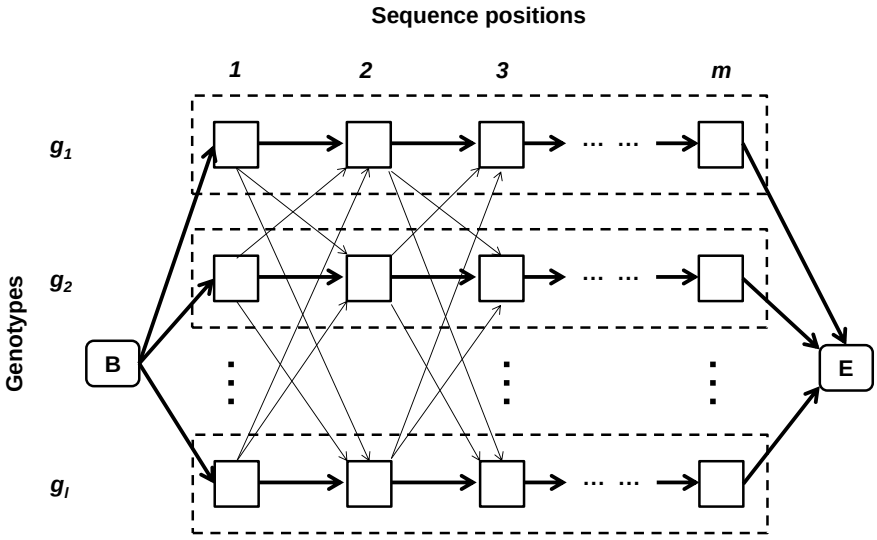

Figure 3.1: Simplified topology of the jumping profile hidden Markov model. The genotypes $g_1, \ldots, g_l$ correspond to hidden states of a hidden Markov model. Thick arrows indicate intragenotype transitions whereas thin arrows indicate transitions between genotypes. For simplicity insertions and deletions are not visualized. Figure is based on Schultz et al. (2006).

(low transition probabilities) to prevent overfitting (high frequency of genotype changes). This penalty parameter was optimized using cross validation on synthetic data sets. The prediction of genotypes and recombination breakpoints is then carried out by computing the Viterbi path, which provides the most probable path of hidden states through all sequence positions (see Section 2.5.2). Note that such a path assigns one particular state and, thus, exactly one genotype to each sequence position of the input sequence. The forward-backward algorithm for hidden Markov models, which computes the position-specific posterior of the hidden state variable, was employed to assign confidence values to the genotype predictions (Schultz et al., 2009).

3.2 Genotyping HBV Dual Infections *in vitro*

Clonal analysis of the viral quasispecies is the widely accepted gold-standard method to identify and genotype dual infections. Amplicons, derived from primers designed to bind to all known HBV genotypes, are cloned in a plasmid vector and transfected into *Escherichia coli*. After the bacteria are cultivated, a number of individual bacteria are picked from the Petri dish to sequence the contained DNA fragments. The resulting sequences are clonal, represent the composition of the viral quasispecies, and are further analyzed by phylogenetic or other genotyping methods. The number of clones selected and sequenced is a crucial parameter that determines the sensitivity of the analysis. The minority genotype might be overlooked if the number of clones is too low. High numbers of clones increase the lab work and lab costs. A binomial model can be applied to relate the number of clones to the expected sensitivity. In case of varying affinity of the primers to the respective genetic variants present in the sample the binomial model might underestimate the number of required clones.

Other methods to identify intergenotype dual infections, which are more cost-effective than clonal analysis, were developed. For example, Naito et al. (2001) designed genotype-specific primers to extract amplicons of genotype-specific length, which can be separated using gel electrophoresis. Genotyping by the use of DNA fragments, which are complementary to genotype-specific conserved regions within preS1 or HBsAg, are another cost-effective and sensitive method to identify intergenotype dual infections (Kato et al., 2003; Osiowy and Giles, 2003). An enzyme-linked immunosorbent assay (ELISA) for genotyping was developed by Usuda et al. (1999). In this assay five monoclonal antibodies identify the genotype by selective binding to corresponding epitopes on the product of the preS2 region. ELISA is available as a commercial kit, which makes it interesting for large-scale surveys.

3.3 Materials and Methods

3.3.1 Data Likelihood of Single and Dual Infections

We now briefly introduce the dual infection model, a statistical model to identify and genotype intra- and intergenotype dual infections using population-based sequence data. Let p_{gj} denote PGSNDs for the HBV genotypes $g \in \{\mathrm{A}, \ldots, \mathrm{H}\}$ and sequence positions j in a given multiple sequence alignment $\mathcal{A}$ of a set of reference sequences. Let $\mathcal{M}$ be

the set of all model alternatives, which are the *single infection models* $\{\mathrm{A}, \ldots, \mathrm{H}\}$ that correspond to the HBV genotypes and the *dual infection models* $\{\mathrm{A\text{-}A}, \mathrm{A\text{-}B}, \ldots, \mathrm{H\text{-}H}\}$ that consist of all unordered pairwise combinations of the single genotype models. For each model $M \in \mathcal{M}$ the position-wise data likelihood of an input sequence represented by $r_i, \ldots, r_{i+n}$ needs to be specified. Each r_j corresponds to one sequence position j with respect to the alignment $\mathcal{A}$ and comprises either a single DNA base ($r_j \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$) or an International Union of Pure and Applied Chemistry (IUPAC) ambiguity code of DNA bases ($r_j \in \{\mathrm{K}, \mathrm{M}, \mathrm{R}, \mathrm{S}, \mathrm{W}, \mathrm{Y}, \mathrm{B}, \mathrm{D}, \mathrm{H}, \mathrm{V}, \mathrm{N}\}$). The ambiguity codes, extracted from population-based sequence data, indicate the presence of multiple genetic variants at a single genome position.

Single infection likelihood. For the single genotype models, the likelihood of an unambiguous sequence position $r_j = b$ is computed by

$$P(r_j = b|g) = p_{gj}[b].$$

As introduced before, $p_{gj}[b]$ denotes the fraction of base b being observed at position j for genotype g and therefore is used to compute the likelihood of base b given genotype g.

For an ambiguous sequence position $r_j = (b_1, b_2)$ where b_1 and b_2 are two divergent bases the likelihood of the sequence position is computed as

$$P(r_j = (b_1, b_2)|g) = p_{gj}[b_1] \cdot p_{gj}[b_2].$$

Ambiguous positions for a single genotype model can be explained by the concept of viral quasispecies. Some strains within a patient might exhibit mutations at certain positions. Nevertheless, all strains originate from the same initial viral variant with which the patient was infected.

Dual infection likelihood. A dual infection is manifested by a combination of two, not necessarily different genotypes $g, h \in \{\mathrm{A}, \ldots, \mathrm{H}\}$. The likelihood of the input sequence is computed using the assumption that population-based sequencing provides the union of the bases of the two underlying viral strains as shown in Figure 3.2. Thus, if the sequence position r_j is unambiguous, there is only a single base b at position j in both strains. This case is not unusual since the overall intergroup divergence of genotypes is in the range between 8% and 17% (Arauz-Ruiz et al., 2002). The likelihood of an unambiguous sequence position is computed as

$$P(r_j = b|g, h) = p_{gj}[b] \cdot p_{hj}[b].$$

The case of an ambiguous sequence position of cardinality two $r_j = (b_1, b_2)$ is difficult to interpret because the phase information of the two bases is lost in the sequencing process, i.e. we do not know to which genotype, g or h, each of the two bases, b_1 and b_2, belong. This corresponds to two biological events: in the first, b_1 is present in a strain of genotype g and b_2 is present in a strain of genotype h, in the second, b_1 is present in a strain of genotype h and b_2 is present in a strain of genotype g. The likelihood of the first event is given by $p_{gj}[b_1] \cdot p_{hj}[b_2]$, while the likelihood of the second event is $p_{gj}[b_2] \cdot p_{hj}[b_1]$. As these events are mutually exclusive, the likelihood of an ambiguous sequence position is obtained by summation:

$$P(r_j = (b_1, b_2)|g, h) = p_{gj}[b_1] \cdot p_{hj}[b_2] + p_{gj}[b_2] \cdot p_{hj}[b_1].$$

Sequence position	1	2	3	4	5	6	7	8	9	10	11	12	13
Strain 1	A	T	**C**	A	A	C	G	**A**	A	T	G	**G**	A
Strain 2	A	T	**T**	A	A	C	G	**G**	A	T	G	**A**	A
Population-based sequencing read	A	T	**CT**	A	A	C	G	**AG**	A	T	G	**AG**	A

Figure 3.2: Dual infection with two viral strains both of which occur at frequencies above the detection threshold of the sequencing technology. Population-based sequencing provides the position-wise union of the two, not necessarily distinct bases of the two viral strains. This gives rise to sequencing ambiguities wherever the viral strains differ.

Regularization. The likelihood computation is extended by a regularization procedure that accounts for sequencing errors in the input sequence and inaccuracies in the PGSNDs. The adjusted position-wise likelihood P^* of a sequence position given any model $M \in \mathcal{M}$ is computed as

$$P^*(r_j|M) = p_{\text{error}} \cdot k_j + (1 - p_{\text{error}}) \cdot P(r_j|M),$$

where k_j equals 1/4 if $r_j \in \{\text{A}, \text{C}, \text{G}, \text{T}\}$ is unambiguous and 1/6 if $r_j \in \{\text{K}, \text{M}, \text{R}, \text{Y}, \text{S}, \text{W}\}$ is ambiguous with cardinality two. Ambiguous sequence positions with cardinality greater than two are not included in the likelihood computation. The error constant p_{error} is set to 1%, the reported error rate of Sanger sequencing (Keith et al., 1993). Sequence positions known to be associated with drug resistance (see Table 2.3) are excluded from the analysis as they might bias the evaluation.

Maximum likelihood principle. The complete input sequence $r_i, \ldots, r_{i+n}$ is classified to be either a single infection or a dual infection of the respective genotype(s) using the adjusted total likelihood $P^*(r_i, \ldots, r_{i+n}|M)$, $M \in \mathcal{M}$ that comprises all adjusted position-wise likelihoods:

$$P^*(r_i, \ldots, r_{i+n}|M) = \prod_{j=i}^{i+n} P^*(r_j|M). \tag{3.1}$$

Equation (3.1) together with the maximum likelihood principle facilitate the analysis of short sequence fragments but fail in situations in which recombinant forms are involved. To analyze recombinant forms it is not sufficient to compute $P^*(r_i, \ldots, r_{i+n}|M)$ as different single or dual infection models need to be applied for different sequence positions.

Detection of recombinants. The jumping profile hidden Markov model was applied in combination with the dual infection model to address complex dual infections with at least one recombinant sequence involved. The resulting model, which again is a HMM, is visualized in Figure 3.3. The single and dual infections models $M \in \mathcal{M}$ correspond to the set of hidden states of the HMM. The input sequence $r_i, \ldots, r_{i+n}$ corresponds to the

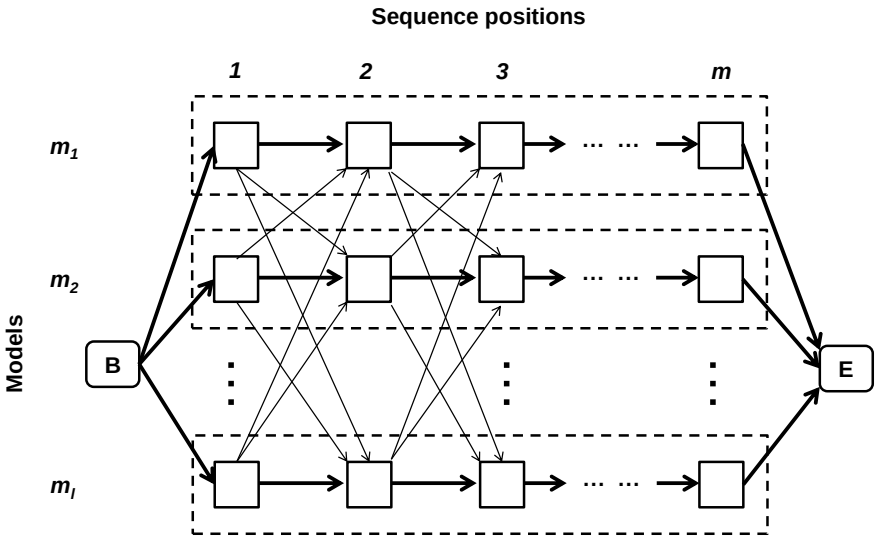

Figure 3.3: Topology of the combination of the dual infection model with the jumping profile hidden Markov model. Each model $m_1, \ldots, m_l$ refers to either a single infection model $\{\mathrm{A}, \ldots, \mathrm{H}\}$ that correspond to the HBV genotypes or to a dual infection model $\{\text{A-A}, \text{A-B}, \ldots, \text{H-H}\}$ that correspond to unordered pairwise combinations of HBV genotypes. The models $m_1, \ldots, m_l$ define position-wise probability distributions that emit HBV sequence data. Thick arrows indicate intra-model transitions whereas thin arrows indicate transitions between models. Figure is based on Schultz et al. (2006).

sequence of observed symbols. Emission probabilities $e_M(r_j)$ depending on the model M and the sequence position j are given by the adjusted position-wise likelihoods:

$$e_M(r_j) = P^*(r_j|M).$$

Transition probabilities $a_{M_k M_l}$ from model $M_k \in \mathcal{M}$ to model $M_l \in \mathcal{M}$ depend on the regularization parameter p_{jump}:

$$a_{M_k M_l} = \begin{cases} p_{\text{jump}} & \text{if } M_k \neq M_l \\ 1 - (|\mathcal{M}| - 1) \cdot p_{\text{jump}} & \text{if } M_k = M_l. \end{cases}$$

The regularization parameter p_{jump} prevents frequent model changes and was set to 10^{-7} (according to Schultz et al. (2012)). Inference in the HMM is performed by computing the most probable sequence of hidden states using the Viterbi algorithm (Section 2.5.2). Confidence of predictions are visualized using the position-specific posterior distributions of the hidden states, which were computed using the forward-backward algorithm (Section 2.5.2).

3.3.2 Training and Test Data

GenBank Data-set

The dual infection model utilizes PGSNDs derived from genotype-annotated sequence data. 3796 HBV full genome sequences were downloaded from GenBank (Benson et al., 2012). The sequence meta-information was parsed to annotate the genotype. Genotypes I and J were excluded from the data set as only very few of these sequences are available. Sequences without genotype annotation were excluded along with sequences whose annotations included the keywords "defective", "non-functional", or "recomb". Additionally, the sequences with accession numbers AB486012, AF461362, AY293309, EF103284, EU305546, EU939634, GQ377573, GU357844, and HQ231877 to HQ231885 were removed from the data set due to evidence from the NCBI genotyping web-service implicating recombinants or false genotype annotations. Duplicate sequences were also removed. The goal of the filtering steps was to obtain a data set of high quality, non-redundant, and non-recombinant sequences with correct genotype annotation that could be used to train and to validate our genotyping methodology. This procedure resulted in 1791 genotype-annotated sequences (270 for genotype A, 339 for genotype B, 636 for genotype C, 304 for genotype D, 159 for genotype E, 54 for genotype F, 12 for genotype G, and 17 for genotype H). Subgenotype predictions were based on sequence data and subgenotype annotations supplied in the literature (Norder et al., 2004; Tallo et al., 2008). Pairwise alignments of the 1791 GenBank sequences with the reference strain AM282986 form the alignment $\mathcal{A}$. PGSNDs were obtained by counting the frequencies of Adenine, Cytosine, Guanine, and Thymine per reference sequence position and genotype.

Synthetic Test Data

Synthetic test data was generated based on sequence data of known genotype from the GenBank data set. Dual infections were simulated by constructing the position-wise union of sequence bases after alignment to the reference strain AM282986. This position-wise union of two sequences emulates the perception of how population-based sequence data of patients with HBV dual infections arise, when both strains are above the sequencing detection sensitivity (see Figure 3.2). This procedure generated sequence data with high numbers of ambiguities representing dual infections that simulate population-based sequence data of *in vivo* dual infections. In total three test sets (TS1 to TS3) were generated. TS1 contains single infections and synthetic intra- and intergenotype dual infections for the genomic regions TP, RT, RN, HBsAg, X, and Core (see Table 2.1 for a description of the genomic regions). TS2 consist of HBsAg sequences only. This part of the genome is sequenced during routine diagnostics and enables the evaluation of the dual infection model in a clinical setting. The sequences in TS2 were impaired by random noise. TS3 consists of full genome sequence data and addresses the ability of the model to genotype complex dual infection including recombinant forms. We now detail the generation of the three test sets.

TS1. TS1 consists of all 1791 GenBank sequences, which we assume represent single infections, and 3600 synthetic intra- and intergenotype dual infections per genomic region (TP, RT, RN, HBsAg, X, and Core). For each of the six genomic regions and for each

combination of genotypes A to H, 100 test sequences were created by randomly selecting two sequences of the respective genotypes. The procedure generated 3600 (36 combinations times 100 samples) test sequences, of which 800 were intragenotype (e.g. A and A) and 2800 were intergenotype (e.g. A and B) for each genomic region.

TS2. The impact of sequencing errors on the identification and genotyping of dual infections was analyzed. TS2 was created sampling randomly 100 single infection HBsAg sequences, 100 intragenotype dual infection HBsAg sequences, and 100 intergenotype dual infection HBsAg sequences from TS1. Each sequence was modified by either adding (for single infection sequences) or removing (for intra- and intergenotype dual infection sequences) ambiguities to simulate sequencing errors. For each sequence, random ambiguities at rates of 1%, 2%,..., 70% were added or removed at random positions. This in total gave rise to 3 times 100 times 70 equals 21000 test sequences in TS2.

TS3. TS3 consists of full genome test sequences. This test set addresses the ability of the model to genotype complex dual infections including recombinant sequences. 1000 test sequences were generated by combining three full genome sequences of distinct genotype from the GenBank data set. To account for an uneven distribution of the genotypes in the GenBank data set, three distinct genotypes were sampled with equal probability. Then, three sequences corresponding to the three distinct genotypes were sampled. Two of the three sequences were used to build a recombinant sequence with three recombination breakpoints each. The three recombination sites were randomly placed in the genome with only one limitation: the recombination breakpoints had to be at least 300 bases apart from each other and from the beginning and the end of the genome. The random placement ensures that recombination breakpoints are placed in different genes at different positions and cover highly variable and highly conserved regions of the HBV genome. The recombinant sequence was then combined with the third sequence to simulate dual infections. The process of generating TS3 is visualized in Figure 3.4.

Performance assessment. Computational experiments based on the GenBank sequence data (TS1 - TS3) were performed with five-fold cross-validation. The PGSNDs were computed on the respective training sequences and the test sequences were generated using the respective test folds of the data sets. Accuracy, sensitivity, and specificity were used to assess prediction quality.

Genotype references. Reference sequences X02763, D00329, X01587, V01460, X75657, X75658, AF160501, and AY090460 were used for HBV genotypes A to H (Arauz-Ruiz et al., 2002; Stuyver et al., 2001).

Patient Test Data

HBV sequences from routine diagnostics performed between 2002 and 2010 at the Institute of Virology (University of Cologne) were employed. The HBV genome region encoding partly the polymerase and HBsAg were amplified and sequenced according to Schildgen et al. (2004).

Clonal analysis. The oligonucleotides HBrt-AccIII-fw (GTCACTTCCGGAGACTACTGTTGTTAGACGACG) and HBrt-RsrII-rev (GCGCATCGGTCCGGCAGATGAGAAGGC) were used to amplify the HBV polymerase of patient isolates from the original serum samples. The amplification products and the vector pCH9-3091 were cut with AccIII and

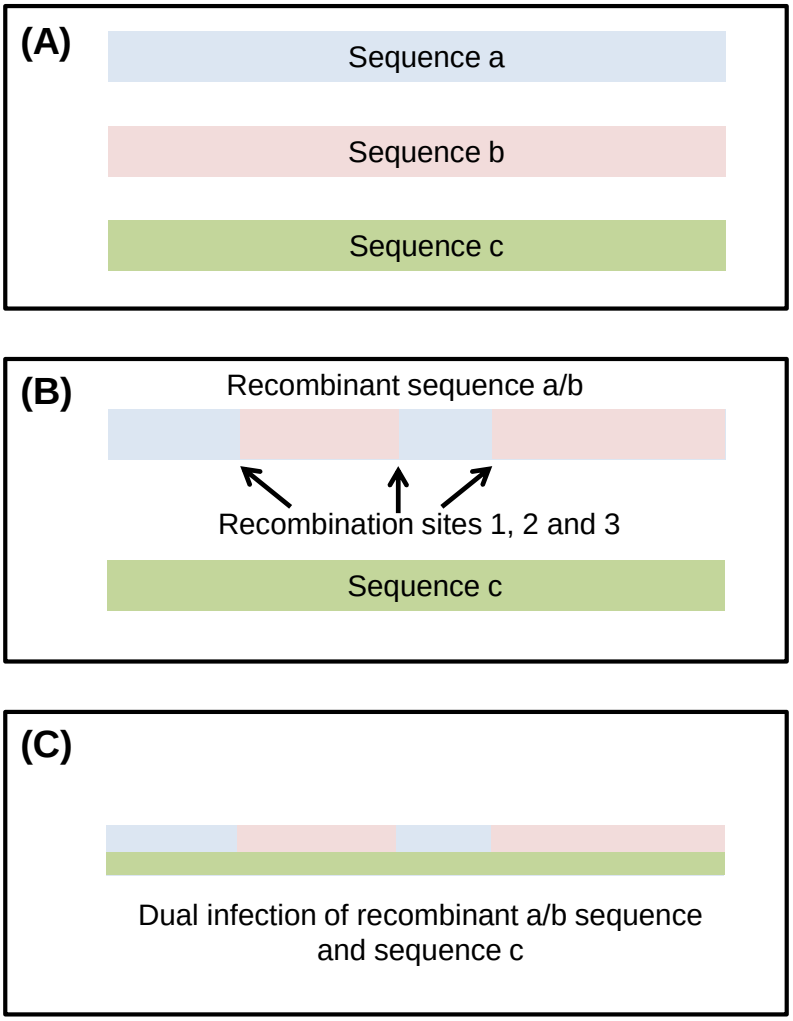

Figure 3.4: The figure visualizes the generation of synthetic test data in test set TS3. (A) Three full genome sequences of distinct genotype are randomly selected from the GenBank data set. (B) Sequence a and b are combined to a recombinant sequence a/b with three randomly placed recombination sites. (C) A dual infection of recombinant sequence a/b and sequence c is simulated by building the positions-wise union of the respective bases according to Figure 3.2.

Type of Infection	Measure	TP	RT	RN	HBsAg	X	Core
Single infection	Accuracy	99.8%	100.0%	99.8%	100.0%	99.8%	99.6%
Intergenotype DI	Accuracy	99.4%	100.0%	99.7%	100.0%	98.8%	98.5%
Intragenotype DI	Accuracy	47.8%	48.6%	47.5%	42.2%	44.7%	45.0%
Intragenotype DI	Accuracy*	99.9%	100.0%	100.0%	100.0%	99.1%	99.5%

Table 3.1: Prediction performance on TS1. TS1 consists of single infection sequences and synthetic intra- and intergenotype dual infections. The measure Accuracy* considers an intragenotype dual infections to be correctly genotypes if either the intragenotype dual infections with the correct genotype is identified or if a single infection with the correct genotype is predicted. The genomic regions TP, RT, RN, HBsAg, X, and Core are detailed in Table 2.1.

RsrII preceding a ligation step. After transformation using JM109 (Promega) clones were picked and analyzed by sequencing. All sequencing was performed using oligonucleotides HBsAg-KpnI-fw (gtcactggtaccatggagagcacaacatcaggattc), HBsAg-MHL-fw (gtcactcatatgctcttcatcctgctgctatgcc), and HBV S6antisense (cKttgacaDactttccaatcaatag). The products were sequenced with the ABI PRISMTM 3130xl and edited with DNASTAR Lasergene. The lab work to derive the clonal sequences was carried out by Maria Neumann-Fraune at the Institute of Virology (University of Cologne). The resulting sequences were analyzed using the NCBI genotyping and the jumping profile hidden Markov model webservices (Rozanov et al., 2004; Schultz et al., 2009).

3.4 Evaluation based on Synthetic Data

The dual infection model was derived from position- and genotype-specific nucleotide distributions, which were computed using full genome sequence data from GenBank annotated with genotype. The dual infection model was applied to three data sets of synthetic data (TS1 to TS3) and to clinical patient sequences to identify and genotype HBV dual infections.

Performance on TS1. Synthetic inter- and intragenotype dual infections were created by randomly combining pairs of sequences of the respective genomic regions. Depending on the genomic region, between 98.5% to 100% of the intergenotype dual infections were correctly identified including the correct combination of genotypes (Table 3.1). Intragenotype dual infections were correctly identified as such by the dual infection model in 42.2.% to 48.6% of the cases. Nevertheless, for 99.1% to 100% of the intragenotype infections the correct genotype was identified, which means that only the intragenotype dual infection could not be inferred. Additionally, the dual infection model genotyped 99.6% to 100% of the GenBank data set correctly.

Performance on TS2. Model predictions were very robust with respect to randomly added or removed ambiguities. Randomly added ambiguities in the range of 1% to 32% did not lead to any falsely positive predicted intergenotype dual infection for single infection sequences (Figure 3.5A). False positive intragenotype dual infections were observed depending on the level of noise. For example, 15% ambiguities resulted in 14 (14.0%) out

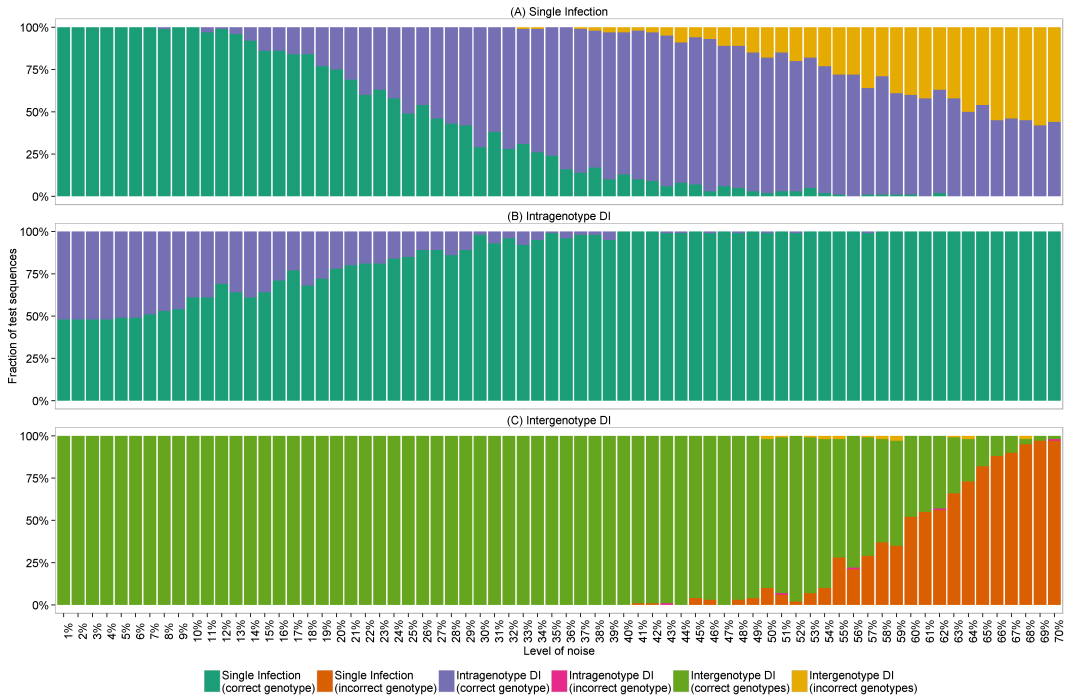

Figure 3.5: Prediction performance on TS2. TS2 consists of (A) single infection sequences, (B) intragenotype dual infection sequences, and (C) intergenotype dual infection sequences that were altered by increasing level of noise displayed on the x-axis.

of 100 false positive intragenotype predictions and 30% ambiguities resulted in 71 (71.0%) out of 100 false positive intragenotype predictions. Adding more than 32% ambiguities led to rare false positive intergenotype dual infection predictions.

Randomly removing 1% to 10% of the ambiguities of intragenotype dual infection sequences did not decrease notably the number of correctly genotyped sequences (Figure 3.5B). Nevertheless, removing more than 10% of the ambiguities hampered the correct identification of intragenotype dual infections.

The correct genotyping of intergenotype dual infection sequences was also not sensitive to randomly removing ambiguities. Removing up to 40% of the ambiguities did not lead to any false genotyping result (Figure 3.5C).

Performance on TS3. The utilization of the jumping profile hidden Markov model in combination with the dual infection model facilitated the accurate genotyping of complex dual infections that involved recombinant forms. Figure 3.6 shows a sample solution derived from this approach. Of the 1000 synthetic test sequences 96.4% were genotyped correctly. The accuracy at the position level was 98.4%. 98.0% of the simulated recombination breakpoints were identified correctly (correct model change and recombination site within 100 bases from the true location) with a median deviation from the ground truth location of 5 bases.

Performance of baseline method. Previous authors identified intergenotype dual infections using a dissimilarity score and a set of genotype reference sequences. The se-

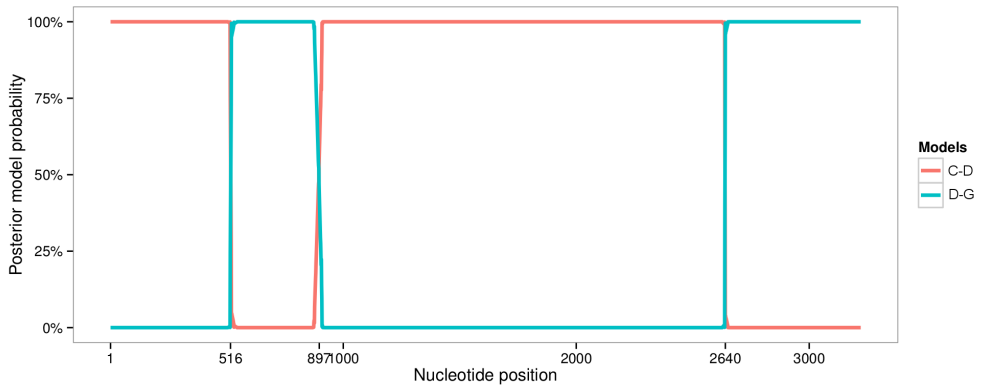

Figure 3.6: The plot shows the position-wise posterior probability of the two models C-D (dual infection of genotypes C and D) and D-G (dual infection of HBV genotypes D and G). This result indicates a dual infection with a genotype D strain and a C/G recombinant strain. Three recombination sites are indicated by the changing posterior probabilities of the two models from approximately zero to approximately one. The ground truth recombination site locations are indicated on the x-axis at positions 516, 897, and 2640. The model estimated recombination sites positions 517, 895, and 2638, respectively. These positions differ from the ground truth in this case by only one, two, and two bases, respectively.

quence dissimilarities of the input sequence to the set of genotype reference sequences were computed and if all dissimilarity scores exceed a cutoff value of 6.0%, the input sequence was considered to represent an intergenotype dual infection (Mallory et al., 2011). We applied this method to our synthetic data set TS1 of synthetic intra- and intergenotype dual infection sequences. The resulting sequence dissimilarities for intragenotype vs. intergenotype dual infections showed distinct distributions but overlapped for each genomic region (Figure 3.7). Thus, the utilization of the dissimilarity score did not facilitate complete separation, and inferred neither intragenotype nor intergenotype dual infections. Additionally, we found that the 6.0% dissimilarity cutoff proposed by Mallory et al. (2011) was too high for optimizing accuracy on TS1. For example, only 61.4% of the HBsAg intergenotype sequences exceeded this cutoff. In contrast, a cutoff of 3.0% sequence dissimilarity yielded sensitivity in the identification of intergenotype dual infections of 99.8%, specificity of 84.9% and the accuracy amounted to 92.4%, while a cutoff of 5.0% yielded sensitivity of 83.7%, specificity of 94.6%, and accuracy of 89.2%. Table 3.2 summarizes the performance evaluation of the sequence dissimilarity method for all genomic regions. Suitable cutoffs were found to depend on the genomic region. We also applied the dissimilarity measure to our test set with randomly introduced ambiguities (TS2) and obviously found that, as soon as the percentage of generated ambiguities exceeded the dissimilarity cutoff, all sequences were classified as intergenotype dual infections. These evaluations show that, on our synthetic data set, methods based on the sequence dissimilarity score could not identify intergenotype dual infections as accurately as the dual infection model and easily get hampered in the presence of sequencing noise.

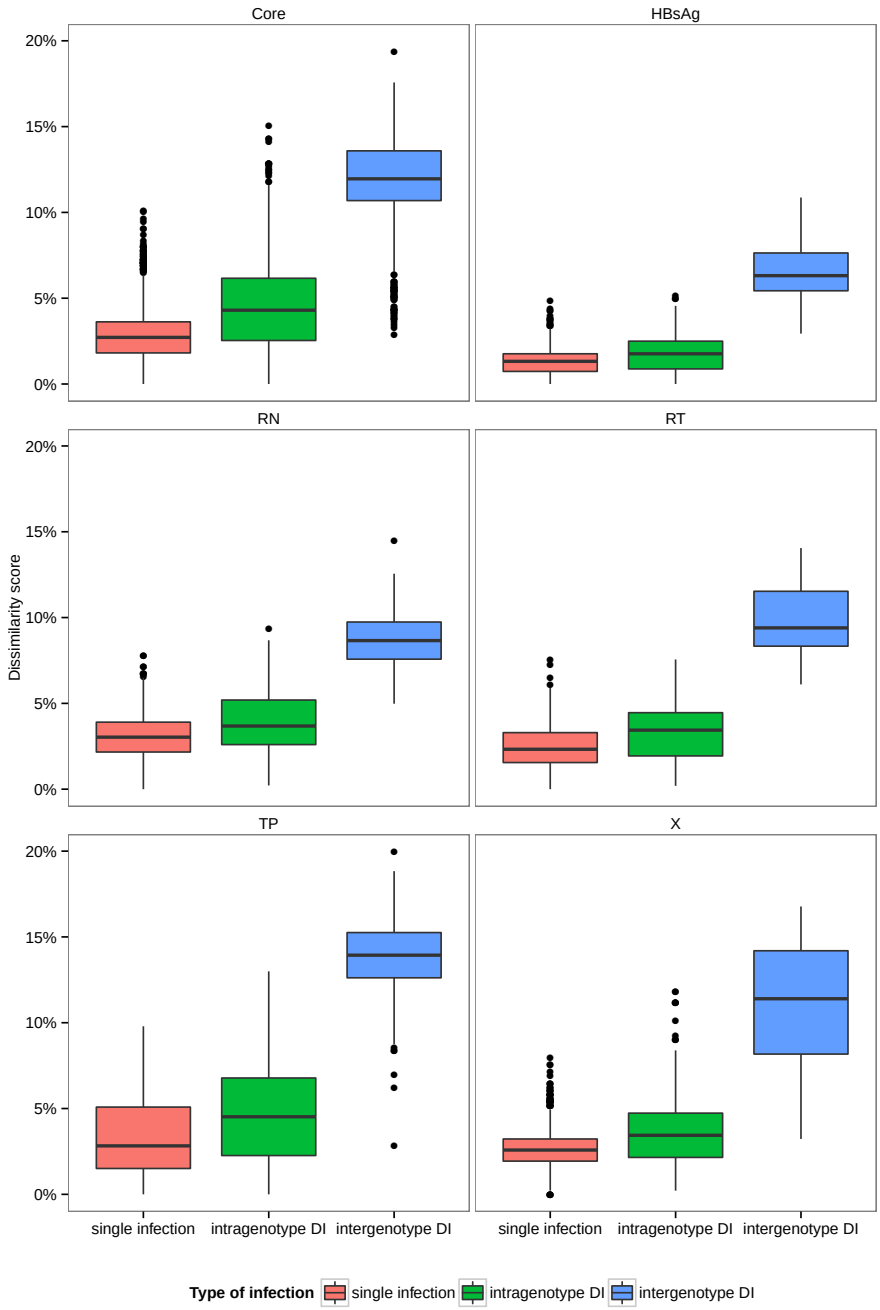

Figure 3.7: Distribution of sequence dissimilarities to the closest genotype reference sequence for single infections (GenBank data set) vs. intragenotype vs. intergenotype dual infections (data set TS1). Sequence dissimilarity scores overlap for all genomic regions TP, RT, RN, HBsAg, X, and Core (see Table 2.1). Thus, the sequence dissimilarities as a baseline method to identify intergenotype dual infections do facilitate separation of intra- and intergenotype dual infections with limited accuracy.

Method	Measure	TP	RT	RN	HBsAg	X	Core
DS 3%	Sensitivity	99.9%	100.0%	100.0%	99.8%	100.0%	99.9%
	Specificity	35.3%	41.3%	32.0%	90.3%	34.8%	31.9%
	Accuracy	69.6%	72.4%	68.1%	95.4%	69.4%	68.0%
DS 4%	Sensitivity	99.9%	100.0%	100.0%	95.5%	99.8%	99.2%
	Specificity	44.9%	63.5%	58.5%	99.0%	62.0%	46.4%
	Accuracy	74.1%	82.9%	80.5%	97.2%	82.1%	74.4%
DS 5%	Sensitivity	99.9%	100.0%	99.9%	84.0%	97.8%	97.9%
	Specificity	53.8%	83.3%	74.6%	99.9%	81.1%	61.2%
	Accuracy	78.3%	92.2%	88.0%	91.5%	89.9%	80.7%
DS 6%	Sensitivity	99.9%	100.0%	96.4%	59.7%	92.7%	95.3%
	Specificity	63.7%	95.3%	90.6%	100.0%	87.3%	73.3%
	Accuracy	82.9%	97.8%	93.7%	78.7%	90.2%	85.0%
DI model	Sensitivity	99.7%	100.0%	100.0%	100.0%	99.9%	99.4%
	Specificity	99.9%	100.0%	100.0%	100.0%	99.0%	99.6%
	Accuracy	99.8%	100.0%	100.0%	100.0%	99.5%	99.5%

Table 3.2: Predictions performance of the sequence dissimilarity score (DS) method with cutoff values of 3%, 4%, 5%, and 6% and the dual infection (DI) model to separate intragenotype vs. intergenotype dual infections on synthetic data set TS1. To allow a direct comparison of the dual infection model with the dissimilarity score method, sensitivity, specificity, and accuracy refers only to the identification of intergenotype dual infections and does not account for the correct identification of the involved genotypes. Thus, accuracy results differ from the values presented in Table 3.1.

3.5 Evaluation based on Patient Data

We employed sequence data obtained from chronically infected patients (n = 241) who were diagnosed within the routine diagnostics at the Institute of Virology (University of Cologne) to evaluate the dual infection model. A population-based sequencing procedure covered the reverse transcriptase domain from amino-acid position 88 (median value) to position 235 (median value) with a median sequence length of 454 bases. Patient sequences were genotyped and screened for dual infections by, first, applying the dual infection model and, second, by applying the baseline method based on sequence dissimilarity scores (Table 3.3). The dual infection model identified eight intergenotype dual infections (one A-D, six A-G, and one D-G) and four intragenotype dual infections (one A-A, two D-D, and one E-E). In contrast, the dissimilarity score method with a 5% dissimilarity cutoff labeled seven samples as intergenotype dual infections, five of which were classified as genotype D single infections by the dual infection model. The 3% dissimilarity cutoff was exceeded by 51 of 241 (21.1%) patient sequences, eleven of which matched with the predictions of the dual infection model.

Dual infection model	Dissimilarity score							
	A	B	C	D	E	F	G	Mixture
A	72	0	0	0	0	0	0	0
A-A	1	0	0	0	0	0	0	0
A-D	0	0	0	0	0	0	0	1
A-G	0	0	0	0	0	0	5	1
B	0	13	0	0	0	0	0	0
C	0	0	6	0	0	0	0	0
D	0	0	0	103	0	0	0	5
D-D	0	0	0	2	0	0	0	0
D-G	0	0	0	0	0	0	1	0
E	0	0	0	0	8	0	0	0
E-E	0	0	0	0	1	0	0	0
F	0	0	0	0	0	1	0	0
G	0	0	0	0	0	0	21	0

Table 3.3: The table contains the counts of patient sequences per genotyping result using the dual infection model (rows) and the sequence dissimilarity score method (columns). For example, as indicated in the first row and the first column, 72 patient sequences were genotypes as a single infection with genotype A by the dual infection model and by the sequence dissimilarity score method. The dual infection model identified eight intergenotype and four intragenotype dual infections. The sequence dissimilarity score method with a 5% dissimilarity cutoff labeled seven patient samples as intergenotype dual infections of which only two coincide with the predictions of the dual infection model.

Validation using sets of clonal sequences. To validate the predictions of the dual infection model, cloning experiments were performed. In three of twelve suspected cases of HBV dual infections a total number of 54 clones were picked, sequenced and genotyped by

the National Center for Biotechnology Information genotyping web-service. In nine cases a shortage of material or insufficient polymerase chain reaction amplification obviated potential experiments. For the first patient sample (A-G dual infection, accession number JQ776529) 21 clones were successfully sequenced of which 19 were of genotype G, one was of genotype A, and one clone was an A/G recombinant. The recombinant sequence was analyzed with the jumping profile hidden Markov model that inferred two breakpoints. The sequence was estimated to be of genotype G between nucleotide positions (with respect to AM282986) 131 and 375 and between 726 and 1053 and of genotype A between nucleotide positions 376 and 725. For the second patient sample (A-G dual-infection, accession number JQ776530) 17 clones were sequenced of which 14 were of genotype G, two were of genotype A, and one was a recombinant of genotypes A and G. The jumping profile hidden Markov model estimated that nucleotide positions 226 to 518 were of genotype A and nucleotide positions 519 to 998 were of genotype G. For the third patient sample (A-G dual infection, accession number JQ776532) a number of 16 clones were successfully sequenced. 13 were of genotype G and three were found to be A/G recombinants. The three recombinant sequences covered nucleotide positions 388 to 1132, 402 to 1171, and 389 to 1093 with estimated breakpoints at nucleotide positions 725, 765, and 782. To summarize, the predictions of the dual infection model were confirmed *in vitro* for these three patient samples as clonal HBV variants of the predicted genotypes were found in the respective patient sera. The three original patient sequences did not exceed the 5% dissimilarity cutoff and were falsely classified as genotype G by the use of the dissimilarity score method.

Intragenotype dual infections. Three of the four putative intragenotype dual infections showed evidence of multiple subgenotypes. Predictions were further investigated by a follow-up analysis in which subgenotype-specific nucleotide distributions were employed by the dual infection model. The A-A dual infection sequence (JQ776527) was identified as an A1-A2 intersubgenotype dual infection. The sequence contained 19 of 454 (4.2%) ambiguities at the nucleotide level which implies the presence of two subgenotypes. Further, its HBsAg sequence contained the ambiguities sS207SN and sL209LV which are characteristic for the presence of subgenotypes A1 and A2. A1 usually encodes Asparagine at position 207 and Leucine at position 209 and A2 usually encodes Serine at position 207 and Valine at position 209 (Norder et al., 2004). Both D-D dual infections were classified as D2-D3 intersubgenotype dual infections. The first of these two sequences (HM174233) contained two different serotypes ayw2 and ayw3 as Arginine was encoded at position 122, Lysine at position 160, Glycine at position 159, and Threonine at position 140 in addition to the HBsAg ambiguity sP127PT (Purdy et al., 2007). In this sequence, Alanine was present at position 128, while the other sequence (HM174239) expressed Alanine and Valine at position 128 and Threonine at position 127 of HBsAg. In both patient samples, Alanine and Threonine were present at position 118 and Methionine and Threonine were expressed at position 125. Thus, the HBsAg amino-acid sequences were in 100% accordance with the conserved amino-acid residues of a combination of the subgenotypes D2 and D3 (Tallo et al., 2008). The sequence HM174233 contained 16 of 394 (4.1%) and the sequence HM174239 contained 13 of 469 (2.8%) ambiguous sequence positions at the nucleotide level.

3.6 Discussion

Taking into account position- and genotype-specific differences in nucleotide distributions can help to identify and genotype HBV dual infections *in silico* from population-based sequencing data. The approach was verified *in silico* with synthetic dual infection data and *in vitro* with clonal experiments of patient samples.

Synthetic data. We showed that the dual infection model has high genotyping accuracy on all genomic regions based on the GenBank sequence data set. This is in concordance with the known fact that HBV facilitates accurate genotyping on all genomic regions due to conserved genetic differences (Myers et al., 2006; Norder et al., 2004). Experiments with synthetic inter- and intragenotype dual infection sequence data showed that intergenotype dual infections can be identified and genotyped with nearly 100% accuracy. Very rare false positive intergenotype dual infections were observed when the model was tested with simulated sequencing errors. HBV genotypes express genotype specific polymorphisms. Consequently, a dual infection with two different genotypes expresses specific combinations of polymorphisms, which are distinct from random noise. Our model based on position- and genotype-specific nucleotide distributions accurately identified these characteristic patterns of ambiguities.

The dual infection model is also capable of identifying intragenotype dual infections. On our synthetic data set these dual infections were identified in 42.2% to 48.6% of the cases depending on the genomic region. The model's ability to distinguish between intragenotype dual infections and sequencing noise was demonstrated by a false positive rate of only 14.0% for intragenotype dual infections in the presence of a very high level of sequencing noise (15% sequence ambiguities).

The dual infection model was combined with the jumping profile hidden Markov model to genotype complex dual infections that involve recombinant sequences. This approach identified 96.4% of the complex synthetic dual infections correctly. The accuracy at the position level was 98.4%. 98.0% of the simulated recombination sites were identified correctly with a median deviation from the ground truth location of 5 bases.

Patient data. The dual infection model identified eight (3.3%) intergenotype and four (1.7%) intragenotype dual infections in our patient cohort (n = 241). The majority of the intergenotype cases were A-G and D-G dual infections. This confirms previous reports that genotype G frequently co-occurs with other genotypes (Kato et al., 2002; Osiowy et al., 2008). Additionally, one A-D dual infection was identified (JQ776528). This sequence was classified as genotype A by standard genotyping web-services (Myers et al., 2006; Rozanov et al., 2004) in which case the risk of interferon treatment failure would have been underestimated due to the overlooked presence of a genotype D strain. Datta et al. (2009) identified and validated five dual infection sequences (four A-D and one A-C) with accession numbers EU275341, EU275342, EU275344, EU275345, and EU275338. These sequences display the same inaccurate classification by standard genotyping methods but are identified correctly by the dual infection model. The NCBI web-service classifies all five as genotype A and HBV STAR classifies the four A-D dual infections as genotype A and the A-C dual infection remains undetermined.

Validation of our predictions with clonal experiments from the original patient material confirmed three out of three predictions. Clonal HBV variants of the predicted genotypes

were found in the respective patient samples. No prediction was falsified. But only few samples could be investigated *in vitro* due to the shortage of original patient material. The analysis of the clonal data using the NCBI genotyping web-service and the jumping profile hidden Markov model for HBV revealed complex quasispecies compositions as two of the three samples contained genotype A variants, genotype G variants, and A/G recombinants. The detailed composition of the viral populations obtained by the clonal experiments could not be elucidated by the use of the dual infection model alone. First, the dual infection model does not account for mixtures of more than two viral strains. Second, the original patient sequences acquired for routine diagnostics were significantly shorter (median length of 454 bases) than the clonal data (median length of 818 bases).

We predicted four cases of intragenotype dual infections within our patient cohort. The intragenotype predictions are inherently difficult to validate and might only reflect the viral quasispecies within the patient. For three of the four predicted cases we have strong evidence of intersubgenotype dual infections. First, the total number of ambiguous sequence positions was high (13, 16 and 19). Second, the observed amino-acids were in concordance with the conserved residues of the combination of subgenotypes predicted by the dual infection model. Compared to the alternative it seems more likely that the observed combinations of amino-acids result from a dual infection with existing subgenotypes. The alternative would be that intra-patient evolution developed the very same combinations of polymorphisms, which are conserved in the subgenotypes. The fourth intragenotype case (E-E dual infection) remains unresolved. It might be a false positive prediction or there might be two distinct genotype E strains present.

Baseline method. The dual infection model outperformed the identification of dual infections based on sequence dissimilarity scores, a baseline method proposed by Mallory et al. (2011). Our model achieved higher accuracy on synthetic data (100% vs. 89.2% with 5% dissimilarity cutoff) and was more robust with respect to sequencing errors (with 10% sequencing noise: 0% false positive intergenotype dual infections vs. 100% false positive intergenotype dual infections). Using the 5% dissimilarity cutoff the baseline method identified (in our patient cohort) only two of the eight intergenotype dual infections reported by the dual infection model. Three validated dual infections did not exceed the 5% dissimilarity cutoff. Lower cutoffs were not suitable for our patient cohort either, due to false positive intergenotype predictions. Other standard genotyping methods do not account for dual infections and, thus bear the risk of leading to suboptimal treatment decisions when present genotypes are overlooked.

Conclusions. The dual infection model provides an *in silico* alternative to labor- and money-intensive clonal experiments. It requires only HBV sequences obtained by routine diagnostics. Alternatives such as second-generation sequencing are able to analyze this problem with higher sensitivity with respect to minor sub-populations, but none of the systems is expected to replace population-based sequencing in the clinical routine within the near future due to cost reasons.

In summary, we developed the first *in silico* genotyping method that can identify and genotype HBV intergenotype dual infections reliably. The method can elucidate the frequency of dual infections in the routine diagnostics and can be helpful in optimizing antiviral therapy. It was integrated into the web-service geno2pheno$_{[\mathrm{HBV}]}$ in the scope of a Bachelor's thesis (Döring, 2011) and is freely available.

4 Linkage Information from Sequencing Chromatograms

I'll fetch what you wish, and I'll fetch more:
Easy it's true, but then easy things weigh more:
It's there already, yet how we might achieve it,
That's the tricky thing, knowing how to seize it.
...
And so it goes on, yesterday and today.
Still buried in the earth, why, there it is:
The earth is the Emperor's, so it's his.

(J. W. von Goethe, Faust Part II, translation by A. S. Kline)

Population-based Sanger sequencing is a cheap and widely used Population-based Sanger sequencing is a cheap and widely used sequencing technology but suffers from low sensitivity regarding the detection of minor variants and from the loss of linkage information. In Chapter 3 we developed a model to identify and to genotype dual infections. In the case of an intra- or intergenotype dual infection we showed that the composition of the viral quasispecies can be elucidated using genotype- and position-specific nucleotide distributions derived from reference sequences annotated with genotype. In this Chapter we show that short-range linkage can be inferred from population-based Sanger sequencing data in a general scenario. While this might seem promising, we will see that the linkage information available from Sanger sequencing data is limited. Our method has significant restrictions with respect to the distance over which linkage can be inferred and with respect to accuracy. However, short-range linkage is of special interest. In the case of two ambiguous sequence positions within the same codon (Figure 4.1), the amino-acids present in the mixture cannot be determined precisely due to the lack of linkage information. Such short-range linkage information is of clinical relevance as treatment decisions, for instance for human immunodeficiency virus and hepatitis B patients, are often based on population-based Sanger sequencing data of the viral genome (Lengauer and Sing, 2006; Lengauer et al., 2007; Zoulim et al., 2009).

Notable attempts have been made to infer linkage information from Sanger sequencing data of the human genome (Dmitriev and Rakitov, 2008; Flot et al., 2006; Flot, 2007; Seroussi and Seroussi, 2007; Sousa Santos et al., 2005). These methods are tailored to genomes that contain a mixture of two different sequences (diploid genomes), one of which harbors an insertion or deletion. The chromatogram downstream of the insertion or deletion shows a high number of double peaks as the two genetic variants are superimposed with a phase shift. These methods rely on additional information, i.e. a reference sequence, a set of possible single nucleotide polymorphisms (SNPs), or the availability of both the forward and the reverse chromatogram. Otherwise, these methods can only be applied in situations in which the two mixture sequences are sufficiently similar and the

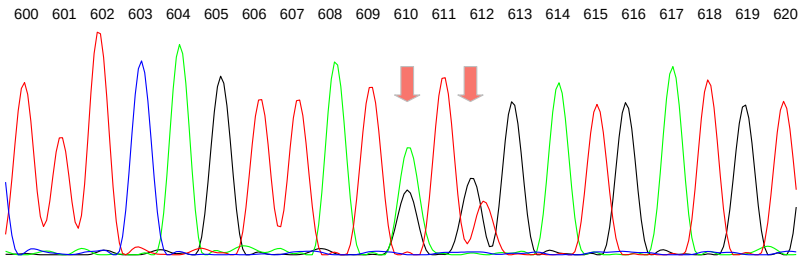

Figure 4.1: The sequencing chromatogram shows two nearby ambiguous sequence positions 610 and 612. At position 610 adenine and guanine are present. At position 612 adenine and thymine are present. Positions are numbered with respect to the reverse transcriptase of the hepatitis B virus genome. This chromatogram raises the question which of the bases at positions 610 and 612 are present in the same clonal variant.

analyzed fragment is significantly longer than the insertion or deletion (Dmitriev and Rakitov, 2008). The method we present here does not rely on diploid genomes that contain a heterozygous insertion or deletion. It is tailored to viral genomes that can have a more complex quasispecies without insertions or deletions. Additionally, we explore a source of information that is encoded in the peak heights of the sequencing chromatograms and has not been considered so far.

Modern Sanger sequencing chromatograms exhibit heterogeneous peak heights that result from different affinities of the polymerase for dideoxynucleotides rather than natural nucleotides during PCR amplification (Lee et al., 1992; Li et al., 1999; Tabor and Richardson, 1995). The rate of incorporation has been found to also depend on the up- and downstream subsequence (Kwok et al., 1994; Parker et al., 1996). This effect is referred to as *sequence context-dependent incorporation of dideoxynucleotides*. Some methods for estimating the frequencies of two alleles present at an ambiguous sequence position account for this effect by comparing the observed peak heights to the peak heights of reference chromatograms (Carr et al., 2009; Kwok et al., 1994). We explore this effect to infer short-range linkage information.

The key assumption of our method is that the collection of peak heights of a mixture of clonal variants (which are called haplotypes, in the following) is the proportion-weighted mixture of the peak heights of the underlying haplotypes. The underlying physico-chemical reasoning is as follows: context-dependent incorporation of dideoxynucleotides resulting in context-dependent peak heights occurs at the level of each polymerase molecule that incorporates dideoxynucleotides with an affinity depending on the DNA subsequence it is processing. The intensity level of the fluorescence observed in the sequencing chromatogram is the sum of all individual molecular fluorescence impulses. Thus, knowing the peak heights of all possible haplotypes present in a mixture, given that these show distinct patterns, renders the determination of the haplotype composition feasible. As context-dependent incorporation of dideoxynucleotides is a local effect, our methodology will only infer linkage information over a short genomic region, maybe up to five bases. A Gaussian noise model was added to the mixture assumption to account for variation in peak heights. Marginal

model likelihoods in combination with the posterior marginals of the model parameters are used for model selection.

This Chapter is structured as follows. First, we describe related work, methods to estimate relative allele frequencies based on sequencing chromatograms. In this context we present a preliminary analysis that has motivated our work (Section 4.2). In Section 4.3 we lay out the methodical details and in Section 4.4 we analyze prediction results based on *in silico* and *in vitro* data. We conclude the Chapter with a discussion (Section 4.5). Note that this Chapter elaborates on work initially presented in Beggel et al. (2013b). The study was performed in cooperation with Maria Neumann-Fraune from the University of Cologne who helped with the design of the study and performed the laboratory work.

4.1 Quantifying Allele Frequencies

Sanger sequencing chromatograms facilitate the estimation of the relative allele frequencies at ambiguous sequence positions (Carr et al., 2009; Kwok et al., 1994; Manion et al., 2009; Qiu et al., 2003). The reasoning is straightforward: given an ambiguous sequence position (double peak) each peak height gives the amount of fluorescently tagged dideoxynucleotides that was incorporated at this position. Thus, the quotient of the peak heights reflects the relative allele frequencies of the two bases at the ambiguous site in the DNA substratum, assumed that there are no biases. Three types of biases hamper the direct interpretation of the peak heights as allele frequencies:

- During the sequencing amplification (which essentially is a PCR as discussed in Section 2.4.1), the relative frequencies of different DNA variants in the sequencing substratum might not be conserved. The chromatogram can only reflect the mixture fractions of the sequencing substratum after the amplification step, which might not reflect the true mixture fractions of the original sample. The parameters of the PCR should be adapted to minimize such distortions as discussed in Carr et al. (2009).

- Different amounts of natural nucleotides and dideoxynucleotides for the four bases contained in the reaction tube lead to altering peak heights. The peak heights need to be normalized using other peaks of the corresponding base prior to interpretation. A base-specific linear scaling transformation is used for this purpose. We refer to these scaling factors as the normalization parameters $\gamma_A, \gamma_C, \gamma_G, \gamma_T$.

- Peak heights observed in chromatograms depend on the sequence context as discussed in the introduction of this Chapter (effect of sequence context-dependent incorporation of dideoxynucleotides). Existing methods to quantify the relative allele frequencies of SNPs based on chromatograms can be divided into two groups: reference-based methods, which account for this effect and try to correct for the corresponding bias, and reference-free methods, which do not.

We now briefly introduce a reference-free method (Qiu et al., 2003) and a reference-based method (Carr et al., 2009). Given a sequencing chromatogram of a mixture of at least two genomic variants that has at least one ambiguous sequence position with two alleles present (we refer to these as base 1 and base 2), let p_1 and p_2 denote estimates for

the relative frequencies of the two alleles at the ambiguous position under analysis. Qiu et al. (2003) compute the base-specific normalization parameters $\gamma_A, \gamma_C, \gamma_G, \gamma_T$ such that the base-specific average height of all peaks in the chromatogram is 1000. Then, p_1 is computed as:

$$p_1 = \frac{h_1}{h_1 + h_2}. \tag{4.1}$$

Here h_1 and h_2 are the peak heights of base 1 and base 2 after normalization. Equation (4.1) ensures that $p_1 + p_2 = 1$ if p_2 is computed accordingly. Thus, equation (4.1) provides consistent estimates for the two allele frequencies p_1 and p_2.

Carr et al. (2009) compute the normalization parameters locally using up to four neighboring positions of the base under inspection. Then, to compute p_1, a reference chromatogram is required that has base 1 at the ambiguous position and the same sequence as the mixture otherwise. The peak height of base 1 in the mixture chromatogram is compared to the peak height of base 1 in the reference chromatogram. Thus,

$$p_1 = \frac{h_1}{h_1^*}. \tag{4.2}$$

Here h_1 and h_1^* refer to the peak heights of base 1 in the mixture and in the reference chromatogram, respectively. p_2 is computed as $p_2 = h_2/h_2^{**}$, where h_2^{**} is the peak height of base 2 in a reference chromatogram that has base 2 at the ambiguous position and is sequence identical to the mixture otherwise. Note that this may provide estimates p_1 and p_2 that do not need to fulfill $p_1 + p_2 = 1$. Carr et al. (2009) do not explain how divergent estimates p_1 and p_2 may be interpreted.

Reference-based methods have several drawbacks compared to reference-free methods. First and foremost, these methods require at least one reference template to determine the unambiguous peak height. Second, it is unclear how to apply reference-based methods in the context of several nearby ambiguous positions. Third, reference-based methods may provide inconsistent estimates p_1 and p_2 with $p_1 + p_2 \neq 1$. Reference-free methods do not face these difficulties and, due to their simplicity, a reference-free method for quantification of ambiguous positions was integrated into a commercial sequence editing software (Manion et al., 2009). On the other hand, reference-free methods might be not as accurate as reference-based methods as they do not account for sequence context-dependent incorporation of dideoxynucleotides. Despite major improvements in the 1990s with respect to the sequencing chemistry (see Section 2.4.1) which have led to quite balanced chromatograms, peak heights might still differ by up to 20%. This is reflected in limited quantification accuracy of reference-free methods, which strongly depend on the genomic position being analyzed.

4.2 Preliminary Analysis

In a preliminary study we analyzed a set of sequencing chromatograms from three dilution series published in Carr et al. (2009). This motivated the mixture assumption and provided evidence that sequencing chromatograms contain linkage information and that reconstructing the haplotype composition might be feasible. Each dilution series was created by mixing different pairs of DNA fragments that differ at a single nucleotide position.

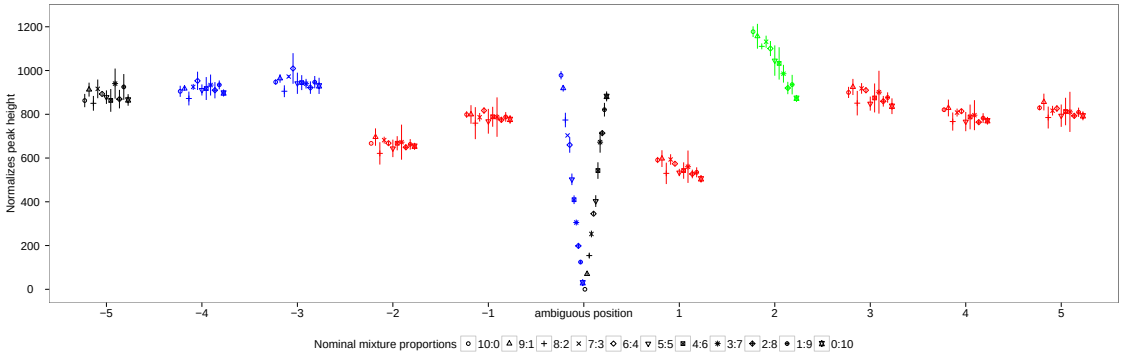

Figure 4.2: The figure shows the median normalized peak heights of the chromatograms of a dilution series (dilution series 1, obtained from Carr et al. (2009)) sorted by nominal mixture proportion. The normalized peak heights before the ambiguous sequence position are almost identical for all nominal mixture proportions. At the ambiguous sequence position and at two to four bases downstream of the ambiguous sequence position a smooth and apparently linear transition between the peak heights of the samples with nominal mixture proportions 10:0 and 0:10 can be observed. The colors indicate the bases: Adenine is green, Cytosine is blue, Guanine is black, and Thymine is red.

Experimental mixture proportions were 10:0, 9:1, 8:2, 7:3, 6:4, 5:5, 4:6, 3:7, 2:8, 1:9, and 0:10. Thus, eleven samples per dilution series were available. Each sample was amplified and sequenced three times.

In the analysis we were interested in changes of peak heights near the ambiguous sequence positions and found that the normalized peak heights upstream of the ambiguous sequence positions were almost identical for all nominal mixture proportions. At the ambiguous sequence position and at the two to four bases downstream we observed a smooth and almost linear transition of the peak heights between the samples with nominal mixture proportions 10:0 and 0:10 (Figure 4.2). This gave rise to the mixture assumption: the peak heights of a mixture of haplotypes are the proportion-weighted mixtures of the peak heights of the underlying haplotypes. The mixture assumption is commonly accepted for the ambiguous sequence position in a dilution series. Reference-based methods for quantifying allele frequencies rely on this principle (Carr et al., 2009; Kwok et al., 1994). We found that the hypothesis is true for all peak heights. The mixture assumption becomes evident near ambiguous sequence positions where the peak heights of the underlying haplotypes differ due to the effect of sequence context-dependent incorporation of dideoxynucleotides. Figure 4.2 shows that there are informative dependencies between peak heights at different positions within sequencing chromatograms, e.g. given the peak height at positions 2, the relative frequency of the bases present at the ambiguous sequence position can be estimated. This source of information might be used to infer linkage information from sequencing chromatograms.

Haplotype	Position 610	Position 612	Position 614	Comment
v_1	Adenine	Guanine	Adenine	Wild-type
v_2	Guanine	Guanine	Adenine	rtM204V
v_3	Adenine	Thymine	Adenine	rtM204I
v_4	Guanine	Thymine	Adenine	rtM204V
v_5	Adenine	Guanine	Thymine	Wild-type
v_6	Guanine	Guanine	Thymine	rtM204V

Table 4.1: Six clonal hepatitis B virus genomes (haplotypes) were created by varying three nearby nucleotide positions. Haplotypes v_1, v_2, v_3, v_4 express all possible combinations of two variants Adenine and Guanine present at position 610 and two variants Guanine and Thymine present at position 612. These nucleotide variants are clinically relevant as they represent two primary resistance mutations rtM204I and rtM204V (see Section 2.3.2). Haplotypes v_1, v_2, v_5, v_6 express all possible combinations of two variants Adenine and Guanine present at position 610 and two variants Adenine and Thymine present at position 614. The mutation at position 614 is clinically not relevant. It was introduced to test our approach with two ambiguities that are three bases apart from each other. Positions are numbered with respect to the reverse transcriptase domain.

4.3 Materials and Methods

4.3.1 *In vitro* Experimental Setup

To evaluate the mixture assumption we designed and created two *in vitro* test sets, each consisting of four clonal HBV variants that differ at two nucleotide positions along with two sets of *in vitro* mixtures of the four respective haplotypes.

The wild-type HBV plasmid vector pCH9-3091 was modified by site-directed mutagenesis. This resulted in a set of six clonal HBV variants (Table 4.1). The clonal variants v_1, v_2, v_3, v_4 represent all possible haplotypes for the two ambiguous positions 610 and 612 in the reverse transcriptase domain shown in Figure 4.1. Likewise, the haplotypes v_1, v_2, v_5, v_6 represent all possible haplotypes for two SNPs at positions 610 and 614. Test set 1 (TS1) consists of 29 *in vitro* mixtures of the four haplotypes v_1, v_2, v_3, v_4 and test set 2 (TS2) consists of 42 *in vitro* mixtures of the four haplotypes v_1, v_2, v_5, v_6. Mixtures were prepared by mixing equal amounts of DNA of the respective haplotypes. The *in vitro* mixtures were submitted to independent sequencing reactions, each with the same set of oligonucleotides.

Clonal variants or mixtures of clonal variants were amplified and sequenced using two distinct established protocols and sequencing machines. TS1 and TS2 were prepared according to Schildgen et al. (2004) and Zhang et al. (2007), respectively. Sequencing of TS1 was performed on ABI 3130xl using BigDye version 1.1 while TS2 was sequenced on ABI 3730 using BigDye version 3.1. The use of two different sequencing protocols was due to interim technological development and not originally intended in the design stage of the study. However, incorporating two technologies enables us to document the robustness of our approach with respect to sequencing protocols. The use of standard sequencing proto-

cols, which are also used in routine diagnostics at the Institute of Virology, University of Cologne, allows us to evaluate our method in a setting close to the clinical routine. Nevertheless, as these standard sequencing protocols are not optimized to preserve mixture fractions, we focus on predicting mixture components.

4.3.2 *In silico* Test Data

An *in silico* test set of 1771 samples was created using all possible mixture fractions on a grid with precision 0.05 using the set of haplotypes v_1, v_2, v_3, v_4. Test chromatograms were computed using the mixture assumption according to equation (4.4).

4.3.3 Data Likelihood Computation

The computational approach requires repetitive sequencing of all possible clonal variants in the mixture as training data. Each haplotype has its own characteristic sequencing chromatogram profile due to the effect of context-dependent incorporation of dideoxynucleotides. Then, the chromatogram of a mixture of haplotypes is the proportion-weighted combination of these haplotype-specific profiles (with some variance). The haplotype-specific profiles are similar to the reference chromatograms used in reference-based methods to quantify relative allele frequencies. Haplotype-specific profiles are derived from several sequencing repetitions of a single template to average out noise and obtain more accurate reference peak heights.

In a preprocessing step all sequencing chromatograms were normalized using a sequence region (called the normalization region) upstream from the region under analysis. Each DNA base (A, C, G, T) requires its own normalization parameter. These four parameters were set such that the average peak height of each base in the normalization region is 1000. Let c_{jki} denote the normalized peak heights of the sequencing chromatograms of the four haplotypes with $j = 1, ..., 4$ indexing a set of haplotypes (either v_1, v_2, v_3, v_4 or v_1, v_2, v_5, v_6), $k = 1, ..., q$ indicating the sequencing replicate, and $i = 1, ..., n$ representing all nonzero peak heights in a window of 20 bases around the ambiguous positions. This includes all positions, both ambiguous and non-ambiguous. Thus, $n = 22$. Note that i does not index sequence positions, i indexes nonzero peak heights. The haplotype-specific profiles p_{ji} were computed as the median normalized peak heights of the sequencing replicates:

$$p_{ji} = \text{median}(c_{j1i}, ..., c_{jqi}). \tag{4.3}$$

Given the fractions $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ (with $\alpha_j \geq 0$ and $\sum_j \alpha_j = 1$) of a mixture of four haplotypes the corresponding mixture profile m_i was computed in terms of the weighted sum of the peak heights of the individual haplotype-specific profiles. Here the weights are the fractions of the haplotypes:

$$m_i = \sum_j \alpha_j \cdot p_{ji}. \tag{4.4}$$

The likelihood of an observed chromatogram with peak heights $D = \{h_i, i = 1, ..., n\}$ given a mixture profile m_i is specified using the following conditional distribution assumption: given normalization parameters $\gamma_A, \gamma_C, \gamma_G, \gamma_T$ for each DNA base, the mixture components $M \subseteq \{1, ..., 4\}$, and the corresponding mixture fractions α_j, the normalized

peak heights $\gamma_{B[h_i]} \cdot h_i$ are assumed to be normally distributed with mean equal to the mixture profile peak heights m_i and constant variance σ^2:

$$P(\gamma_{B[h_i]} \cdot h_i | M, \alpha) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(\gamma_{B[h_i]} \cdot h_i - m_i)^2}{2\sigma^2}}. \tag{4.5}$$

Here $B[h_i]$ denotes the DNA base of peak height h_i. The variance σ^2 reflects sequencing-dependent variations in peak heights, which were estimated using sequencing replicates of clonal variants.

In order to apply equation (4.5) (to compute the data likelihood $P(D|M, \alpha)$) the normalization constants $\gamma_A, \gamma_C, \gamma_G, \gamma_T$ need to be estimated. Therefore, the data D is partitioned into the distinct union of $D = D_{\text{eval}} \dot{\cup} D_{\text{fit}}$. D_{eval} contains all ambiguous and D_{fit} the remaining peak heights. $\gamma_A, \gamma_C, \gamma_G, \gamma_T$ are estimated on D_{fit} using the maximum likelihood principle given the conditional distribution assumption (4.5) and the mixture profile m_i. Fitting each normalization constant is an ordinary least-square regression problem for each mixture profile m_i. The likelihood of the observation $P(D|M, \alpha)$ is then computed using equation (4.6). Thus, we evaluate $P(D|M, \alpha)$ by computing $P(D_{\text{eval}}|M, \alpha, D_{\text{fit}})$. After the normalization constants $\gamma_A, \gamma_C, \gamma_G, \gamma_T$ have been estimated using D_{fit}, the peak heights of D_{eval} can be assumed to be independent given the conditional distribution assumption. This allows the likelihood function to be factorized:

$$\begin{aligned} P(D|M, \alpha) &= P(D_{\text{eval}}|M, \alpha, D_{\text{fit}}) \\ &= \prod_{h \in D_{\text{eval}}} P(\gamma_{B[h]} \cdot h | M, \alpha). \end{aligned} \tag{4.6}$$

4.3.4 Model Selection

The primary goal is to determine the haplotypes $M \subseteq \{1, ..., 4\}$ represented by a given chromatogram. Maximum likelihood estimates derived from the conditional distribution assumption overfit and likely produce point estimates indicating that all four haplotypes are present. In order to guard against overfitting regularization is employed by computing the marginal model likelihoods with uniform priors $P(\alpha|M)$ according to the principals of probabilistic reasoning as introduced in Section 2.5.1. Given a model M the marginal model likelihoods are computed as:

$$\begin{aligned} P(D|M) &= P(D_{\text{eval}}|M, D_{\text{fit}}) \\ &= \int P(D_{\text{eval}}|M, \alpha, D_{\text{fit}}) \cdot P(\alpha|M)\, d\alpha. \end{aligned} \tag{4.7}$$

For the setting with two ambiguous sequence positions seven alternative models need to be considered, which are M_{1+4}, M_{2+3}, M_{1+2+3}, M_{1+2+4}, M_{1+3+4}, M_{2+3+4}, and $M_{1+2+3+4}$. Thus, inferring the haplotype composition is a 7-fold classification problem. The notation M_{1+4} is used to indicate a mixture of haplotypes 1 and 4, M_{1+2+3} is used to indicate a mixture of haplotypes 1, 2 and 3, etc.. Other subsets of $\{1, ..., 4\}$, e.g. $\{1, 2\}$, do not exhibit two ambiguous positions in the chromatogram and therefore can be excluded.

To improve model selection, the modes $\beta_1, \ldots, \beta_4$ of the posterior marginals of the model parameters $\alpha_1, \ldots, \alpha_4$ of the full model $M_{1+2+3+4}$ are computed. For each $j \in \{1, 2, 3, 4\}$ let $\alpha_{\bar{j}}$ denote all α_i with $i \neq j$.

$$\begin{aligned}\beta_j &= \arg\max_{\alpha_j} P(\alpha_j|D, M_{1+2+3+4}) \\ &= \arg\max_{\alpha_j} \int_{[0,1]^3} P(\alpha|D, M_{1+2+3+4})\ d\alpha_{\bar{j}} \\ &= \arg\max_{\alpha_j} \int_{[0,1]^3} P(\alpha|M_{1+2+3+4}) \cdot P(D|M_{1+2+3+4}, \alpha)\ d\alpha_{\bar{j}}\end{aligned} \tag{4.8}$$

A model $M \subseteq \{1, ..., 4\}$ is considered to be *inconsistent* with $\beta_1, \ldots, \beta_4$ given a certain cutoff value t if there exists $j \in M$ with $\beta_j \leq t$ and (in case $M \neq M_{1+2+3+4}$) if there is $k \notin M$ with $\beta_k > t$. In words, there is a haplotype j indicated by the model with low β_j and a haplotype k not present in the model with high β_k, which together indicate a strong contradiction of the model with the parameter marginals. The latter condition that haplotype k is not present in the model even though $\beta_k > t$ can and need only to be fulfilled if the model under consideration is not the full model ($M_{1+2+3+4}$). Models that are inconsistent with the modes of the parameter marginals are excluded from the model selection procedure. We refer to this as the application of the *consistency rule*. The cutoff t used to evaluate the prediction performance on TS1 was chosen to optimize accuracy on TS2 and vice versa.

The integrals (4.7) and (4.8) were approximated using a grid of α values of precision 0.025. This resulted in 12341 parameter configurations, for which the log-likelihoods were computed and summed up to compute the marginal model likelihoods and the posterior marginal parameter distributions.

4.3.5 Performance Evaluation

Haplotype reconstruction for two ambiguous sequence positions is a 7-fold classification problem. Prediction performance at the model level is measured in terms of accuracy. Prediction performance is also evaluated at the clonal level. For this purpose, each 7-fold classification problem is interpreted as four 2-fold classification problems defined by the prediction of the presence or the absence of each of the four possible haplotypes. Confidence of predictions is expressed in terms of an uncertainty cutoff. Each prediction is either correct, incorrect, or unassigned. A sample remains unassigned if the marginal likelihood of the best model divided by the marginal likelihoods of all other models falls below the uncertainty cutoff.

4.4 Results

4.4.1 Application to Dilution Series

The mixture assumption was employed to estimate the mixture fractions of three dilution series obtained from Carr et al. (2009). The maximum likelihood estimates for the mixture fractions derived from the chromatograms were close to the nominal mixture fractions for dilution series 1 and 2 with average absolute errors of 0.01 and 0.03, respectively (Figures 4.3A and 4.3B). Dilution series 3 exhibited a nonlinear relationship between nominal and estimated proportions, resulting in an average absolute error of 0.14 (Figure 4.3C). This

nonlinear relationship is likely caused by varying amplification efficiencies of the different clonal variants. Almost identical results for these data were reported by Carr et al. (2009). Note that the variance in the estimates originates from repetitive amplification and sequencing of the same *in vitro* mixture. The experiment described above was repeated while blanking out the ambiguous positions. Thus, we employed only sequence positions -2, -1, 1, and 2 (see Figure 4.2) to estimate the mixture fractions. This resulted in fraction estimates with higher average errors of 0.05, 0.08, and 0.26 and higher variances (Figures 4.3D-4.3E). The mixture assumption facilitates the estimation of the fractions of a mixture without using the ambiguous positions. Previous methods that quantify mixture fractions based on sequencing chromatograms only consider ambiguous positions and neglect the information provided by surrounding peaks (Carr et al., 2009; Kwok et al., 1994; Qiu et al., 2003).

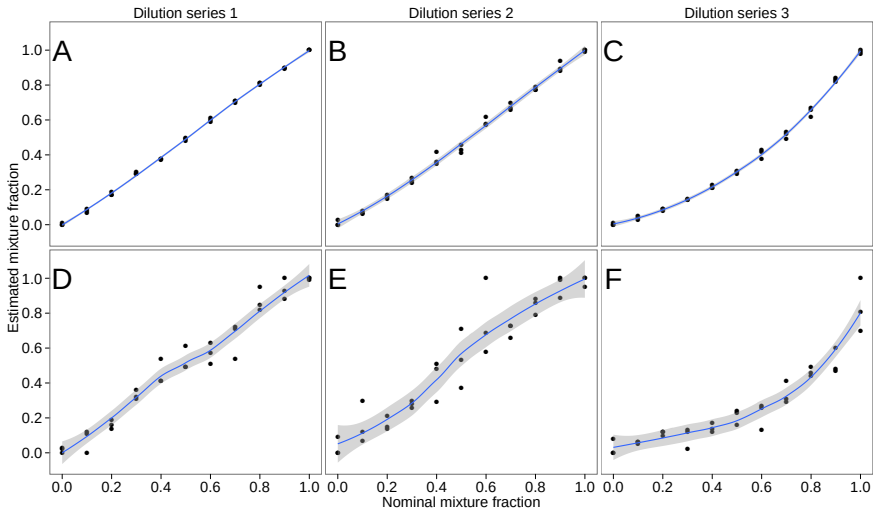

Figure 4.3: All plots show nominal mixture fractions versus estimated mixture fractions for three dilution series. (A-C) uses all peak heights and provides proportions estimates with low error and low variance. (D-F) ignores the peak heights at the ambiguous positions and tries to estimate the mixture proportions only based on the surrounding peak heights of the unambiguous positions. The resulting fraction estimates show higher errors and higher variances. The blue lines are fitted by local polynomial regression lines (LOESS), grey areas indicate 95% confidence intervals.

4.4.2 Complexity of Haplotype Reconstruction

We now focus on the problem of haplotype reconstruction based on sequencing chromatograms in the special case of two ambiguous sequence positions. For this problem four possible haplotypes $1, \ldots, 4$ with fractions $\alpha_1, \ldots, \alpha_4$ need to be considered. Table 4.1 rows 1 to 4 lists the four possible haplotypes for the combination of two important HBV drug resistance mutations within the reverse transcriptase domain. The haplotype reconstruction problem has three degrees of freedom $\alpha_1, \alpha_2, \alpha_3$, which determine α_4 due

to $\sum_j \alpha_j = 1$. An estimate of the fraction of the first ambiguous position p_1 gives rise to $\alpha_1 + \alpha_3 = p_1$ as haplotype 1 and 3 are wild-type at the first ambiguous position. Similarly, an estimate p_2 of the fraction of the second ambiguous position provides $\alpha_1 + \alpha_2 = p_2$. These frequency estimates can be derived from the chromatograms, thus reducing the complexity to one degree of freedom. This implies that the peak heights of the two ambiguous positions impose strong constraints that limit the space of possible solutions to approximately a straight line in the three-dimensional simplex of all possible solutions. Within this set of possible solutions we face the problem of model selection. Further evidence from the surrounding peaks or slightly different configurations of the four ambiguous peaks are required to localize the correct solution.

4.4.3 *In silico* Experiment

To study elementary properties of our approach we created a set of *in silico* test samples. 1771 artificial chromatograms were created for the set of haplotypes $v_1, \ldots, v_4$ (see Table 4.1) based on equation (4.4) using a grid of α values with grid size 0.05. Test data were generated without adding a noise component. The noise-free setting provides an upper bound on the model performance to be expected on real-world data and simplifies the analysis of model characteristics and limitations. Mixture chromatograms were truncated to nucleotide positions 600 to 620. Positions 610 and 612 were employed to compute the data likelihoods. The value of the standard deviation σ was estimated using the sequencing repetitions of the haplotypes. σ was set to 20.94. As peak heights were normalized to have an average height of 1000, a standard deviation of 20.94 implies variation of peak heights of approximately 2.1%.

Only 867 (49%) of the 1771 *in silico* test samples were predicted correctly. Figure 4.4 visualizes all incorrectly predicted test cases separately for each falsely predicted label. The four surfaces of the simplex correspond to the four 3-mixture models M_{1+2+3}, M_{1+2+4}, M_{1+3+4}, and M_{2+3+4}. Test samples that lie in the interior of the simplex but close to one of its faces display mixtures of all four haplotypes of which one haplotype has low frequency (Figures 4.4A-4.4D). These test samples were regularized towards the surface of the simplex by using the marginal likelihoods for model selection. This reveals the first major restriction of the model in terms of sensitivity, which is that low frequency haplotypes in a mixture of four haplotypes can not reliably be detected.

The test set can be divided into two groups: those samples that contain at least one haplotype with a fraction of no more than 10%, and those for which each present haplotype occurs in a fraction higher than 10%. The samples of the first group were predicted correctly in 383 (32.0%) of 1200 cases. The samples of the second group (all present haplotypes exceed a fraction of 10%) were predicted correctly in 484 (85%) of 571 cases.

Figure 4.4E indicates a second systematic failure, which is that 46 test samples (34 $M_{1+2+3+4}$, 5 M_{1+2+3}, and 7 M_{2+3+4}) were incorrectly predicted as mixtures of haplotypes 1 and 4. All of these samples satisfied the set of linear constraints $\alpha_2 - \alpha_3 \leq 0.05$, $\alpha_2 \leq 0.35$, and $\alpha_3 \leq 0.35$. These misclassifications occur due to model regularization as detailed in Section 4.5. Similarly, 41 test samples (32 $M_{1+2+3+4}$, 5 M_{1+2+4}, and 4 M_{1+3+4}) were falsely predicted as mixtures of haplotypes 2 and 3 (Figure 4.4F). These samples satisfied the constraints $2\alpha_1 + \alpha_2 + \alpha_3 - 1 \leq 0.05$, $\alpha_1 \leq 0.30$, and $\alpha_4 \leq 0.30$. To improve prediction

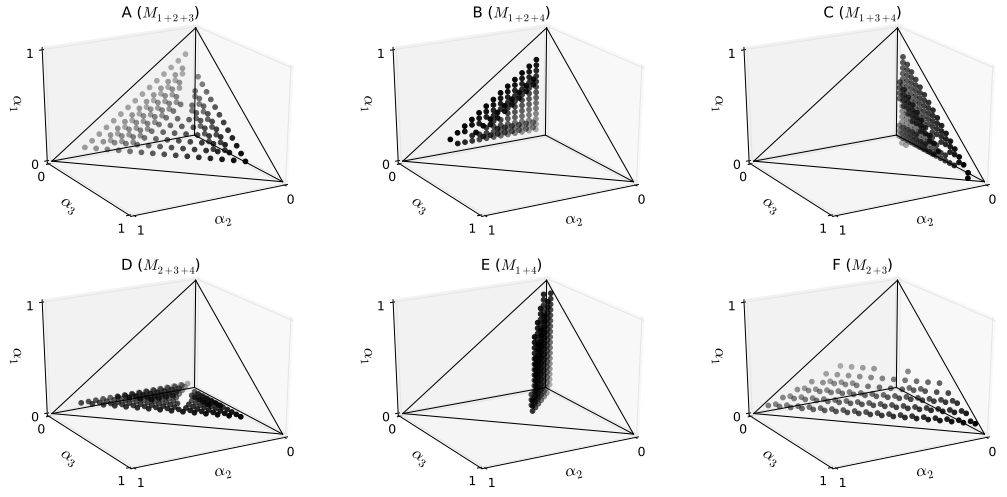

Figure 4.4: 1771 *in silico* test chromatograms were created by computing the mixture profiles on a grid of α values with grid size 0.05. The plots show all falsely classified samples separately for each falsely predicted label. Six major cases of misclassification can be observed. Plots A-D show test samples, which consist of four haplotypes, at least one of which having low proportion. Plots E and F show test samples that were predicted as mixtures of haplotypes 1 and 4 or as mixtures of haplotypes 2 and 3. The data points of plot E satisfy the linear constrains $\alpha_2 - \alpha_3 \leq 0.05$, $\alpha_2 \leq 0.35$ and $\alpha_3 \leq 0.35$ and data points of plot F satisfy $2\alpha_1 + \alpha_2 + \alpha_3 - 1 \leq 0.05$, $\alpha_1 \leq 0.30$ and $\alpha_4 \leq 0.30$, respectively.

performance we analyzed the marginal posterior distributions of the fraction parameters $\alpha_1, \ldots, \alpha_4$. For example, for all of the seven M_{2+3+4} test samples that were misclassified as M_{1+4} we observed that the modes of the posterior marginal distributions of α_1 were zero while the modes of α_3 had a mean of 0.22 ± 0.07. Thus, the modes of the posterior marginals were clearly inconsistent with the model predictions based on the marginal model likelihoods. By checking the consistency of the model predictions with the modes of the posterior marginals using a cutoff of 10% (as described in Section 4.3.4) all misclassified 3-mixture test samples could be corrected. Nevertheless, the consistency rule does not improve prediction performance on the 4-mixture test samples. The inability of the model to distinguish between 2-mixture and 4-mixture test samples (if one of two sets of linear constraints is fulfilled) is the second major limitation of the model in terms of accuracy.

4.4.4 *In vitro* Validation

Our hypothesis was that the haplotype composition of arbitrary mixtures can be reconstructed using information from the ambiguous and the surrounding peak heights. To test this hypothesis, we created two *in vitro* training sets, each composed of six sequencing repetitions of the underlying haplotypes (Table 4.1) to compute the individual haplotype profiles and a respective test set of *in vitro* mixtures. Test set 1 (TS1) based on haplotypes v_1, v_2, v_3, v_4 contained 29 mixtures and test set 2 (TS2) based on haplotypes v_1, v_2, v_5, v_6

contained 42 mixtures. When model selection was solely based on the marginal model likelihoods, prediction accuracy at the model level was 96.6% on TS1 and 47.6% on TS2. By application of the consistency rule as described in Section 4.3.4, prediction accuracy improved to 71.4% on TS2 and remained unchanged on TS1. The respective cutoff used to evaluate the prediction performance on TS1 was chosen to optimize accuracy on TS2 and vice versa. A cutoff of 10% was used for both data sets. Note that model selection in this setting is a 7-fold classification problem and the accuracy expected by chance amounts to only 14.3%. Prediction performance was also evaluated at the clonal level. In this evaluation scheme we treated the 7-fold classification problem as four 2-fold classification problems that amounted to the prediction of the presence or the absence of each of the four possible haplotypes. Using this evaluation we could study the accuracy of the model predictions in more detail compared to the assessment of accuracy in terms of 0-1 error at the model level. For instance, the incorrect classification of a $M_{1+2+3+4}$ test sample as M_{1+2+3} would result in an accuracy of 75% as three of four haplotypes were predicted correctly. Clonal level prediction accuracy was 97.4% for TS1 and 84.5 % for TS2. Figure 4.5 summarizes the prediction performances on TS1 and TS2 both at the model level and at the clonal level. Additionally, we present the results as a function of the prediction confidence expressed by the uncertainty cutoff. For example, at an uncertainty cutoff of 4.0, which means that the marginal likelihood of the best model is at least 4.0 times as high as the marginal likelihood of the second-best model, only 2.6% (5.4%) of the predictions at the clonal level were incorrect, 14.6% (30.4%) were unassigned and 82.8% (64.2%) of the predictions were correct on TS1 (respectively TS2).

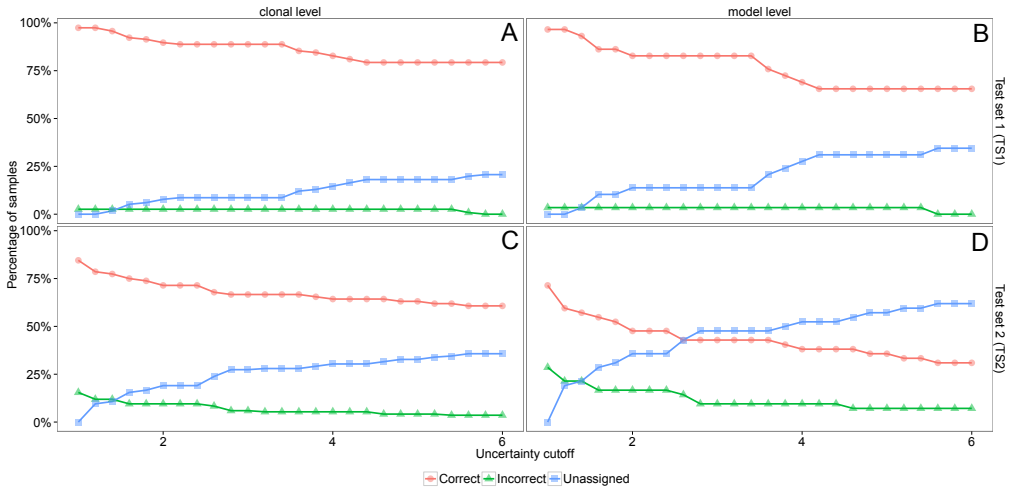

Figure 4.5: The figure shows the prediction accuracy on test sets TS1 (subplots A and B) and TS2 (subplots C and D). Prediction accuracy was evaluated both at the clonal and at the model level. Each test sample is predicted either correctly or incorrectly, or is classified as unassigned. The latter happens when the marginal likelihood of the best model divided by the marginal likelihood of the second-best model falls below the uncertainty cutoff displayed on the x-axis.

4.5 Discussion

We described the first attempt, to our knowledge, to use peak heights from sequencing chromatograms to infer linkage information about nearby ambiguous positions. A generative model describes the expected peak heights of a mixture of haplotypes by combining the chromatograms of the underlying haplotypes. The model was motivated by visual inspection of the peak heights of a dilution series. Our model is a generalization of previous methods for estimating the fractions of one ambiguous sequence position (Carr et al., 2009; Kwok et al., 1994). The main difference is that we employed a generative model, whose parameters are estimated by probabilistic inference rather than by comparing observed peak heights to template peak heights. Previous methods neglected nearby peak heights as they are usually influenced by the ambiguous peaks to be analyzed. In contrast, we exploit this interaction effect to infer linkage information.

Prediction accuracy. The high prediction accuracy of our method on two *in vitro* test sets of 29 and 42 mixture samples shows that sequencing chromatograms do contain linkage information. The prediction accuracy on TS1 was 97.4% at the clonal level and 96.6% at the model level while the accuracy on TS2 was 84.5% at the clonal level and 71.4% at the model level. We found that the posterior marginal distributions of the model parameters α_j are informative of the mixture compositions as prediction performance at the model level increased from 47.6% to 71.4% on TS2. The probabilistic framework provides confidence values for the predictions in terms of posterior distributions, which can be applied to reduce the number of false predictions at the cost of a higher number of unclassified samples.

Model limitations. The evaluation of the model on a set of 1771 *in silico* test samples revealed several limitations of our approach. First, the model cannot reliably detect minor populations with frequency of no more than 10%. In such cases, the predictions are regularized towards simpler models, which do not contain the low-frequency haplotype. Second, we found that if the mixture fractions α_j satisfy one of two sets of linear constraints, the mixture is always predicted to be either a mixture of the haplotypes 1 and 4 or a mixture of haplotypes 2 and 3. The true mixture components in the former case are M_{1+4}, M_{1+2+3}, M_{2+3+4}, and $M_{1+2+3+4}$ and in the latter case are M_{2+3}, M_{1+2+4}, M_{1+3+4}, and $M_{1+2+3+4}$. This becomes plausible by looking at the proportion estimates p_1 and p_2 of the two ambiguous positions. If p_1 and p_2 are equal ($p_1 = p_2$), then the observed peak heights can be interpreted to originate from a mixture of haplotypes 1 and 4 with $\alpha_1 = p_1$ and $\alpha_4 = 1 - p_1$. From $p_1 = p_2$ follows $\alpha_1 + \alpha_3 = \alpha_1 + \alpha_2$ and immediately $\alpha_2 - \alpha_3 = 0$. If additionally α_2 and α_3 have low frequency (below 35%), we obtain the set of linear constraints that are fulfilled by all *in silico* test cases, which were falsely classified as mixtures of haplotypes 1 and 4 (Figure 4.4E). The presence of haplotypes 2 and 3 does not become strongly evident by looking at the surrounding peaks due to their low frequency. Similarly, if the frequency estimates at the two ambiguous positions fulfill $p_1 = 1 - p_2$, the mixture can be interpreted to contain only the haplotypes 2 and 3 with $\alpha_2 = p_2$ and $\alpha_3 = p_1$. From $p_1 = 1 - p_2$ follows $\alpha_1 + \alpha_3 = 1 - \alpha_1 - \alpha_2$ and immediately $2\alpha_1 + \alpha_2 + \alpha_3 - 1 = 0$, which is the constraint satisfied by all *in silico* test cases that were falsely predicted as a mixture of the haplotypes 2 and 3. This observation is reflected in the *in vitro* test set TS2, in which four samples (9.5%) were incorrectly classified at an uncertainty cutoff of 4.0. Three samples were $M_{1+2+3+4}$ mixtures misclassified as M_{2+3} with $2\alpha_1 + \alpha_2 + \alpha_3 - 1$ amounting

to 0.0, 0.0, and 0.025, respectively. Additionally, one M_{1+2+3} sample was misclassified as M_{1+4} with $\alpha_2 - \alpha_3$ equal to 0.025.

To summarize, we think that the misclassifications do not result from inaccurate likelihood computations, but rather from the model regularization. Simpler models are preferred, if the peak heights at the ambiguous positions do not imply a more complex model. This is the case, in particular, when one haplotype in a mixture of four haplotypes has low frequency (misclassified as a mixture of three haplotypes), when $p_1 \approx p_2$ (misclassified as M_{1+4}), or when $p_1 \approx 1 - p_2$ (misclassified as M_{2+3}).

Applicability. We evaluated our hypothesis to perform inference on the haplotype compositions for only two sets of sequence positions within the reverse transcriptase domain of the HBV genome. Nevertheless, peak heights usually display very distinctive patterns. Thus, we assume the approach will also work in other settings. Two different amplification and sequencing protocols in concert with two different sequencing machines were employed to generate the two test sets. This indicates the robustness of the method with respect to sequencing protocols. The reasoning that the problem is essentially one-dimensional after the mixture fractions are estimated implies that the model very likely cannot successfully be extended to infer the haplotype composition in the presence of more than two nearby ambiguities. Additionally, the locality of the effect of sequence context-dependent incorporation of dideoxynucleotides limits the capacity of the approach to infer linkage information of ambiguous positions that are more than a few bases (up to five bases, at most) apart from each other. On test set TS2, in which the two ambiguous positions were three positions apart from each other, prediction accuracy was already impaired in comparison to TS1, in which the two ambiguous positions were only separated by a single base.

Second generation sequencing. The advent of second-generation sequencing technologies has superseded Sanger sequencing in many applications (Mardis, 2008; Metzker, 2010). 2ndGS technologies offer increased sequencing depth and speed through a high degree of parallelization and miniaturization and allow for the detection of minor variants with relative frequencies of as low as 10^{-4} (Gerstung et al., 2012). 2ndGS data also naturally provide linkage information over the whole read length, which may further be extended by the use of paired-end reads (Chaisson et al., 2009). Compared to our approach, 2ndGS technologies are far more sensitive and accurate in determining the haplotype composition of mixtures and can deliver long-range linkage. Nevertheless, Sanger sequencing is cheaper than 2ndGS and 2ndGS will not be accessible in many laboratories for some time to come. Considering this, certain applications, in which the limitations of our approach in terms of sensitivity, accuracy, and limited range are acceptable, may not require 2ndGS.

Conclusions. We have developed and validated an approach to compute the peak heights of a mixture of haplotypes based on the chromatograms of the underlying haplotypes. The model can be used to infer the haplotypes present in a mixture and therefore the short-range linkage for two ambiguous sequence positions. The effect of sequence context-dependent incorporation of dideoxynucleotides, an effect that was previously regarded as detrimental, was employed. The effectiveness of our method shows that short-range linkage information can be inferred from sequencing chromatograms with no further assumptions on the mixture composition. The model also allows the estimation of the fractions of nearby ambiguities and therefore overcomes the limitations of Carr et al. (2009). As a major limitation to its widespread applicability, the model requires the sequencing chromatograms

of all possible haplotypes in the mixture. The source code of our method is available at `http://bioinf.mpi-inf.mpg.de/publications/beggel/linkageinformation.zip`.

5 G-to-A Hypermutation of the HBV Genome

Real biologists who actually do the research will tell you that they almost never find a phenomenon, no matter how odd or irrelevant it looks when they first see it, that doesn't prove to serve a function. The outcome itself may be due to small accidents of evolution.

(E. O. Wilson)

The immune response to infections with the hepatitis B virus is a complex and not well understood system that involves the development of specific antibodies (anti-HBs, anti-HBc, and anti-HBe) and the activation of a broad set of restriction factors within the response of the innate immune system. Understanding this system might lead to a better understanding of the four phases of the natural progression of chronic hepatitis B infections as the phases are mainly determined by the state of the immune system (see Section 2.1.3), and, as a consequence, to new antiviral strategies. An important part of the innate immune system are APOBEC3 (A3) proteins, a subgroup of the APOBEC (apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like) protein family. All members of the APOBEC family are *cytosine deaminases*, which catalyze the hydrolytic deamination of cytosine to uracil (Figure 5.1B). A3 deaminases gained a lot of attention since the gene of the most prominent member of this subgroup, APOBEC3G (A3G), was identified by Sheehy et al. (2002). A3 deaminases have the ability to mutate the HBV genome and the genome of other DNA and RNA viruses. The deamination of cytosine mediated by APOBEC proteins is referred to as *editing*. Chelico et al. (2006) provided a detailed analysis of the molecular mechanisms of the editing activity of A3G, which was shown to bind to single-stranded DNA and to mutate cytosine to uracil while moving from $3'$ to $5'$. During completion of the complementary plus strand, deamination of the (–)-DNA strand results in the translation of uracil into guanine, which becomes visible as G-to-A mutations. Viral genomes with an unexpected high number of G-to-A mutations were first observed in HIV patients (Vartanian et al., 1991; Wain-Hobson et al., 1995). The term *G-to-A hypermutation* is used to describe the phenomenon of up to 80% of guanines being mutated to adenines (G-to-A mutations) in a short sequence region of about 100 bases.

The work described in this Chapter was funded by the Forschungskommission of the Heinrich-Heine-University, Düsseldorf. Project partner Prof. Andreas Erhardt, (formerly Heinrich-Heine-University, Düsseldorf, currently Petrus Hospital, Wuppertal) designed the study and provided patient data. Prof. Carsten Münk, Heinrich-Heine-University, Düsseldorf supported and guided the analysis. The work was initially published in Beggel et al. (2013a). We present a detailed analysis of the G-to-A hypermutation pattern of the complete HBV genome on the basis of second-generation sequencing data of 80 treatment-naïve patients (47 HBeAg-positive and 33 HBeAg-negative). We found that the intensity of G-to-A hypermutation depends on various factors, especially the position along the genome, the HBeAg status, and the state of fibrosis. Additionally, we found that the prevalence

of HBsAg and HBeAg escape mutations are statistically significantly associated with G-to-A hypermutation. We start this Chapter with a short history of the discovery of A3G (Section 5.1.1) and a summary of the APOBEC protein family members (Section 5.1.2). Then, we review the effects of APOBEC proteins on HBV (Section 5.1.3). Section 5.2 presents the methodical background of our study. The results of our analysis are outlined in Section 5.3. In Section 5.4 we conclude with summarizing remarks and a classification of our results with respect to the published knowledge about hypermutation of the HBV genome.

5.1 Biological and Clinical Background

5.1.1 Discovery of APOBEC3G

The study of the APOBEC proteins is closely related to the study of Vif, one of HIV-1's accessory regulatory proteins. HIV-1 belongs to the genus *Lentivirus*, which in turn is part of the family of retroviruses. The genome of retroviruses contains the reading frames Gag, Pol, and Env encoding the viral matrix protein, the viral enzymes, and the surface proteins, respectively. Lentiviruses encode several additional (essential or accessory) regulatory proteins, which in case of HIV-1 are Tat, Rev, Nef, Vif, Vpr, and Vpu. During knock-out experiments performed to study the function of the regulatory proteins, it was observed that Vif-deleted (ΔVif) HIV-1 variants cannot spread in primary CD4 T-cells and macrophages (Gabuzda et al., 1992; Sova and Volsky, 1993; von Schwedler et al., 1993). To be more precise, a collection of cell lines, so-called nonpermissive cells, can be infected with ΔVif HIV-1 variants but produce noninfectious HIV-1 particles. Other cell lines (e.g. Jurkat and SupT1), so-called permissive cells, enable ΔVif HIV-1 to replicate in the absence of Vif. This suggests two possibilities for the role of Vif in HIV-1 replication: either permissive cells express a protein similar to Vif, which can perform the function of Vif in the HIV-1 replication cycle, or Vif counteracts a cellular restriction factor that is only present in nonpermissive cells. A3G was identified to be the counteracting restriction factor in a study conducted by Sheehy et al. (2002), in which two very similar cell lines, one permissive and one nonpermissive, were analyzed using subtractive hybridization techniques. Shortly thereafter, several other members of the APOBEC family were described (Chiu and Greene, 2008).

5.1.2 APOBEC Protein Family

The human APOBEC protein family consists of eleven members: APOBEC1, APOBEC2, activation-induced cytidine deaminase (AID), APOBEC3A-APOBEC3H (A3A-A3H), and APOBEC4 (Jarmuz et al., 2002; Liao et al., 1999; Rogozin et al., 2005). All of these have the ability to edit DNA or RNA by performing the hydrolytic deamination of cytosine to uracil (Figure 5.1B). The seven A3 deaminases (A3A, A3B, A3C, A3DE, A3F, A3G, and A3H) are arranged in a tandem gene array on chromosome 22, which implies that they likely have a common ancestor and arose from successive gene duplication events.

The active sites of the APOBEC proteins, which are involved in DNA and RNA editing, are characterized by a conserved zinc-binding motif His-X-Glu-X_{23-28}-Pro-Cys-X_{2-4}-Cys

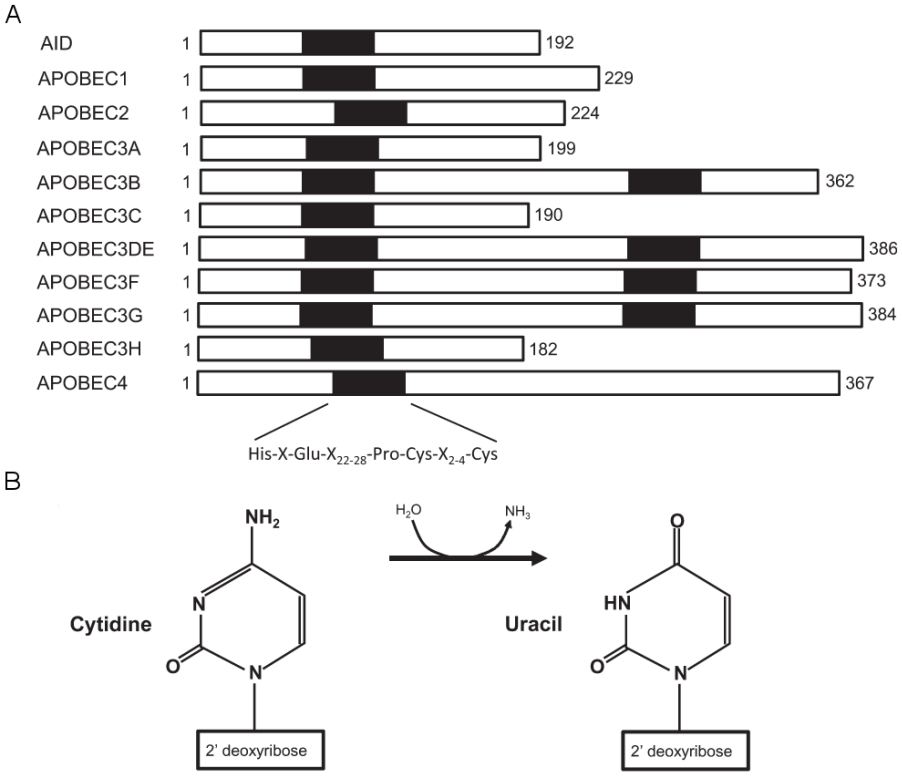

Figure 5.1: (A) Location of the active sites of the members of the APOBEC protein family. The numbers state the first and last amino acids of the cytosine deaminases. The back region indicates the location of the conserved zinc-dependent cytosine deaminase active sites with the amino-acid motif shown at the bottom. (B) Deamination of cytidine (cytosine if incorporated into DNA) to uracil. Permission to reuse this figure was granted by John Wiley & Sons (Janahi and McGarvey, 2013).

(Jarmuz et al., 2002; Wedekind et al., 2003). Of note, A3B, A3DE, A3F, and A3G have two copies of this catalytic site (Figure 5.1A).

The members of the APOBEC family have very distinct biological functions. APOBEC1, AID, and APOBEC3(A-H) are well characterized, while the functions of APOBEC2 and APOBEC4 are not well understood. APOBEC1 mediates the deamination of a specific cytosine at position 6666 in the messenger RNA of apolipoprotein B into uracil-6666, which creates an in-frame stop codon. The truncated (apoB48) and untruncated (apoB100) versions of the apolipoprotein B protein were described to have distinct functions in the lipid metabolism (Mehta et al., 2000; Teng et al., 1993). APOBEC1, in contrast to the other members of the APOBEC family, primarily edits RNA rather than DNA.

The catalytic activities of AID are crucial for the development of mature and diverse antibody responses. AID is expressed in germinal centre B-lymphocytes, where it facilitates cytosine-to-uracil mutations of the immunoglobulin gene. This random deamination of cytosines is referred to as somatic hypermutation of the immunoglobulin gene and aims to increase the diversification of the immunoglobulin gene and facilitates antibody class switches (e.g. from immunoglobulin M to immunoglobulin G) (Muramatsu et al., 1999, 2000).

A3 deaminases are part of the innate immune system and act against exogenous viruses (Delebecque et al., 2006; Yu et al., 2004) and endogenous retroelements (Bogerd et al., 2006; Kinomoto et al., 2007). A3 deaminases inhibit a broad range of viruses, for instance HIV-1, HBV, HCV, human papillomavirus (HPV), herpes simplex virus type 1 (HSV-1), and Epstein-Barr virus (see Vieira and Soares (2013) for a recent review). It is important to note that the antiviral activity of A3 deaminases only partly relies on their ability to edit viral genomes. A3 deaminase-inactive variants were shown to inhibit e.g. HIV-1 with a deaminase-independent mode of action (Bishop et al., 2006, 2008). G-to-A hypermutation of the HCV genome has not been described yet while A3G was shown to be a strong inhibitor of HCV replication *in vitro* (Peng et al., 2011).

Not all members of the A3 subgroup inhibit all exogenous viruses or endogenous retroelements, which can partly be explained by different localizations. A3D, A3F, and A3G are mainly observed in the cytoplasm (Bennett et al., 2008; Kinomoto et al., 2007), A3B is almost exclusively located in the nucleus (Lackey et al., 2012), and A3A, A3C, and A3H are found both in the nucleus and in the cytoplasm (Kinomoto et al., 2007; Li and Emerman, 2011).

5.1.3 APOBEC and HBV

Hypermutated HBV genomes were already observed *in vivo* by Günther et al. (1997), but the molecular mechanisms of hypermutation were unknown. Hypermutated genomes occur *in vivo* at low frequency (below 1%), which makes them difficult to observe and to analyze (Noguchi et al., 2009; Suspène et al., 2005b). Thus, intensive studies could only be performed after the APOBEC genes were identified, which allows specific over- and under-expression of APOBEC genes, and 3D-PCR, a differential DNA amplification technique, became available (Sheehy et al., 2002; Suspène et al., 2005a). 3D-PCR utilizes the fact that AT-rich DNA melts at a lower temperature than GC-rich DNA. This allows hypermutated genomes to be selectively amplified and facilitates the analysis of hypermutated genomes

despite their low frequency.

In vitro. *In vitro* studies showed that the HBV genome is susceptible to editing by APOBEC1 (Gonzalez et al., 2009), AID (Vartanian et al., 2010), and all A3 deaminases except for A3DE (Baumert et al., 2007; Henry et al., 2009; Köck and Blum, 2008; Rösler et al., 2005; Suspène et al., 2005b). HBV's (–)-DNA strand is the preferred editing template resulting in G-to-A hypermutated genomes, which occur in frequencies of 10^{-5} to 10^{-2} depending on the A3 deaminase (Henry et al., 2009). Rare C-to-T hypermutated genomes or G-to-A + C-to-T mixed hypermutated genomes were observed *in vitro* in A3B, A3F, and A3G overexpressed cells, which suggests (+)-DNA strand editing (Suspène et al., 2005b). It is not well understood how A3 proteins can edit HBV's (+)-DNA strand. It might be that the (+)-DNA and (–)-DNA strands temporarily separate, which would enable A3 binding and editing.

Inhibition. Intensive hypermutation of the HBV genome leads to deleterious mutations, which result in non-functional proteins and, obviously, in inhibition of the replication. Furthermore, DNA that has undergone intensive editing (uracil-rich DNA) might be recognized as degraded and might be digested by the combined action of uracil DNA glycosylases and apurinic-apyrimidinic endonuclease (Schröfelbauer et al., 2005; Sova and Volsky, 1993). By the use of deamination-inactive variants it was shown that A3B, A3G, and likely other A3 deaminases inhibit HBV in a deamination-independent mode of action. Deamination-inactive A3G inhibits HBV with the same strength as wild-type A3G, which indicates that the contribution of hypermutation to its inhibitory effect is likely minimal (Noguchi et al., 2007). Subsequently, it was pointed out that A3G is a strong inhibitor of HBV but leads to a relatively low number of edited genomes while A3C, on the other hand, has only little inhibition activity but leads to a high number of mutated genomes (Baumert et al., 2007; Köck and Blum, 2008). A study that focused on A3B showed that the inhibitory effect of deamination-inactive A3B is reduced to approximately 40% compared to its wild-type counterpart (Bonvin and Greeve, 2007). Both of A3B's active sites (see Figure 5.1A) contribute to its antiviral effect, while only the C-terminal active site has deamination activity. The deamination-independent mode of inhibition is not yet fully understood. A3 deaminases are likely incorporated into the HBV capsid and prevent the completion of capsid formation or the reverse transcription of the pregenomic RNA (Nguyen et al., 2007; Rösler et al., 2005).

Dinucleotide context. APOBEC proteins exhibit context-specific editing profiles (Henry et al., 2009; Suspène et al., 2005b; Vartanian et al., 2010). APOBEC1 shows a preference for GA-to-AA mutations and an aversion against GC-to-AC mutations (Gonzalez et al., 2009; Vartanian et al., 2010). AID prefers G-to-A mutations in the GC and GT context, which leads to GC-to-AC and GT-to-AT mutations, respectively (Beale et al., 2004; Vartanian et al., 2010). In contrast to APOBEC1 and AID, all A3 deaminases prefer editing in the GG and GA context. A3G shows strong preference towards GG-to-AG mutations while A3A, A3B, A3C, and A3F have a more balanced dinucleotide preference (Henry et al., 2009; Suspène et al., 2005b). The number of mutations in the GG and GA context vs. the number of mutations in the GC and GT context is used in *in vivo* studies to discriminate between reads edited by APOBEC1/AID and those edited by A3 deaminases (Suspène et al., 2011; Vartanian et al., 2010). Additionally, a very high number of GG-to-AG mutations indicates A3G activity due to A3G's strong preference for mutations

in the GG dinucleotide context. Using the dinucleotide context it was shown that the vast majority of hypermutated genomes observed *in vivo* were edited by A3 deaminases.

Interferon stimulation. Primary human hepatocytes of healthy individuals express low levels of the messenger RNAs of A3B to A3G but the expression of A3B, A3C, A3F, and A3G can be upregulated by interferon stimulation. Upregulation of interferon leads to a significant reduction of HBV replication (Bonvin et al., 2006; Jost et al., 2007; Noguchi et al., 2005; Tanaka et al., 2006). Nevertheless, blocking A3B, A3F, and A3G while upregulating interferon does not abrogate the inhibitory effect of interferon on HBV. Thus, A3B, A3F, and A3G likely contribute only minimally to the antiviral effect of interferon (Jost et al., 2007).

Clinical implications. Despite the importance of A3 deaminases for the innate immune response and the strong antiviral activity of e.g. A3G *in vitro*, relatively little is known about the clinical relevance of A3 deaminases for the natural progression of hepatitis B. Only a small number of studies tried to relate hypermutation to the clinical course of hepatitis B patients. Noguchi et al. (2009) observed that a combined increase of ALT levels and hypermutation levels was associated with a reduction of the viral load. In the same study four cases of spontaneous HBeAg seroconversion were observed after an increase of the number of hypermutated genomes. It is quite natural to link G-to-A hypermutation to HBeAg-negative chronic hepatitis: the most frequent HBeAg escape variant G1896A is a GG-to-AG mutation and GG is the preferred editing dinucleotide of A3G. Several authors hypothesized that A3-mediated hypermutation might play an important role in the natural course of HBV infections (Noguchi et al., 2009; Turelli et al., 2004; Vartanian et al., 2010). For example, A3 deaminases were linked to disease progression towards cirrhosis as these are upregulated in cirrhotic liver tissue from HBV single infected and HBV+HCV dual infected patients (Vartanian et al., 2010).

5.2 Patients and Sequencing Data

We now focus on the methodical details of our study, which aimed to utilize second-generation sequencing data to analyze the phenomenon of G-to-A hypermutation of HBV and to relate hypermutation to fundamental characteristics of patients with chronic hepatitis B.

Patients. The cohort consists of 47 HBeAg-positive and 33 HBeAg-negative treatment-naïve patients. Patients were enrolled in a multicenter, randomized, partially double-blind study that analyzed the effectiveness and safety of pegylated interferon α-2a treatment in HBeAg-positive patients (Lau et al., 2005) and HBeAg-negative patients (Marcellin et al., 2004). Patient characteristics including liver biopsy as well as virological and serological parameters were obtained according to the study protocols. Patient characteristics (detailed in Table 5.1) exhibit statistically significant differences between HBeAg-positive and HBeAg-negative patients with respect to age, ethnicity, HBV genotype, and viral load.

Viral DNA isolation, amplification, and Roche/454 pyrosequencing. Viral DNA was extracted from 200 microliter serum using the QIAamp blood mini-kit. DNA bar-coded primer pairs were designed to generate seven overlapping amplicons (amplicon 1 - 7) covering the whole HBV genome (see Table 5.2). Viral DNA was amplified using a touch-down PCR protocol.

Characteristic	All patients (n=80)	HBeAg-positive patients (n=47)	HBeAg-negative patients (n=33)	P-value
Age, years [mean (range)]	37.2 (18-70)	32 (18-65)	44.6 (20-70)	$< 0.001^{*}$
Male sex [n (%)]	68 (85%)	42 (89.4%)	26 (78.8%)	1.0^{**}
Ethnicity [n (%)]				0.01^{**}
Caucasian	67 (84.8%)	35 (74.5%)	32 (97.0%)	
Other	13 (16.2%)	12 (25.5%)	1 (3.0%)	
HBV genotype [n (%)]				$<0.001^{**}$
A	39 (48.8%)	33 (70.2%)	6 (18.2%)	
D	41 (51.2%)	6 (29.8%)	21 (81.8%)	
ALT [mean (range)]	97 (14.6-300)	98.4 (29.2-300)	94.9 (14.6-300)	0.87^{*}
HBV DNA [median (range)]	9.2 (6.3-14.5)	10.1 (7.7-14.5)	8 (6.3-10.3)	$<0.001^{*}$
Degree of Fibrosis				0.35^{***}
F0	10 (12.5%)	7 (14.9%)	3 (9.0%)	
F1	27 (33.8%)	16 (34.0%)	11 (33.3%)	
F2	28 (35%)	18 (38.3%)	10 (30.3%)	
F3	7 (8.8%)	4 (8.5%)	3 (9.1%)	
F4	8 (10%)	2 (4.3%)	6 (18.2%)	

Table 5.1: The table provides a summary of the patient characteristics. P-values were obtained by * Wilcoxon rank-sum test, ** Fishers exact test, or *** t-test for the respective parameter to be zero in a linear model. ALT levels are provided in terms of international units (IU) per milliliter and HBV DNA levels are provided in terms of copies per milliliter. F0 to F4 indicates the state of fibrosis with F0 and F1 denoting absent or mild fibrosis and F2 to F4 denoting significant fibrosis.

Amplicon	Genomic region
1	57 to 709
2	524 to 1197
3	1093 to 1660
4	1524 to 2113
5	1956 to 2485
6	2297 to 2990
7	2820 to 199

Table 5.2: Sequencing amplicons.

Amplicons were purified using Agencourt CleanSeq beads on a BioMek NX workstation (Beckman Coulter) and quantified fluorometric on a FluoStar Optima (BMG Labtech, Cary, NC, USA). For the emulsion PCRs the amplicons of ten patient samples each were pooled equimolarly. After recovery of the beads and enrichment, approximately 790,000 beads per pool were loaded on one region of a GS FLX PicoTiterPlate and subdivided with a 4-lane gasket. Sequencing was performed on a Genome Sequencer FLX Titanium (Roche-454 Life Sciences).

Sequence data preprocessing. Sequence reads were clipped using phred-equivalent quality scores (with a cutoff value of ten) or clipped after 400 bases at latest. The SMALT program (Wellcome Trust Sanger Institute) was used to map the reads to the AM282986 reference strain. Mapped reads shorter than 100 bases were removed from the data set. In total, 1,360,551 reads were successfully mapped with an average sequence length of 375 $\pm$ 56 bases. The number of reads per patient exhibited a heterogeneous distribution and ranged from 7,047 to 61,623 with a mean of $17,006 \pm 7,332$. The median coverage (number of reads per patient sample and nucleotide position) had a mean of 1,480 $\pm$ 601. In the following we refer to the set of clipped, filtered, and mapped reads as the 454-data.

Hypermutation rates. Several scoring schemes were proposed to identify hypermutated reads. The G-to-A preference was defined as proportion of G-to-A mutations divided by the total number of mutations in the read (Rose and Korber, 2000). The total number of G-to-A mutations was also used as indicator of hypermutation. Mutations were evaluated with respect to the sample consensus sequence based on the 454-data, which included every base with frequency of at least 10%. Each read can be classified as normal or hypermutated based on the number of G-to-A mutations and the G-to-A preference with individual cutoffs. The quotient of the G-to-A hypermutated reads divided by the total number of reads in any genomic region is referred to as the *hypermutation rate* (HMR).

5.3 G-to-A Hypermutation Pattern

HBeAg status. The 454-data were scanned to identify G-to-A hypermutated genomes. Hypermutation rates were computed with a combination of two hypermutation criteria (at least four G-to-A mutations and G-to-A preference of at least 70%) for non-overlapping 100-base windows along the genome (Figure 5.2). Median hypermutation rates per 100-base window varied between 0.0% and 0.047% for HBeAg-positive and between 0.0% and 0.69%

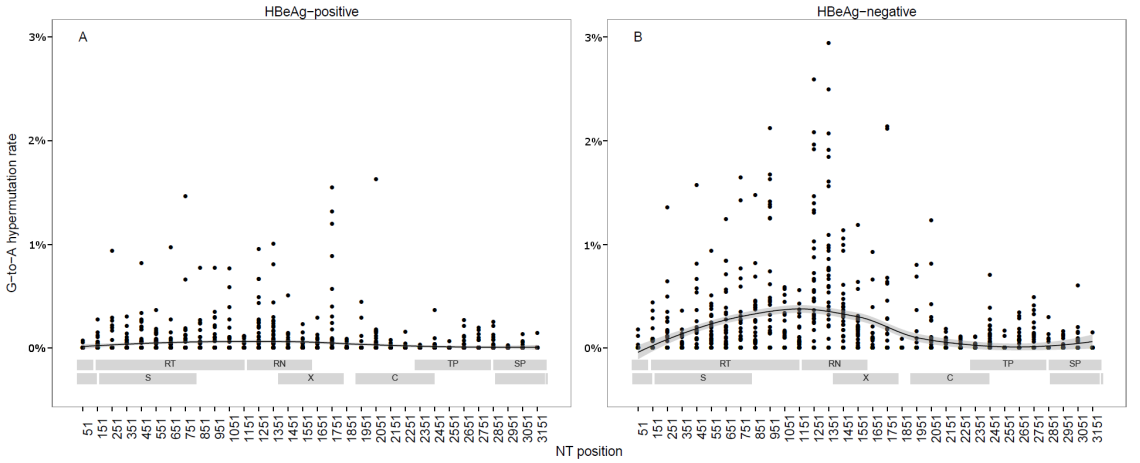

Figure 5.2: Analysis of G-to-A hypermutation rates along the hepatitis B virus genome. The scatter plots show per-patient hypermutation rates across the genome for HBeAg-positive (A) and HBeAg-negative (B) patients. Each point corresponds to one patient and one 100-base window. The hypermutation criterion requires at least four G-to-A exchanges and a G-to-A preference of at least 70% within each 100-base window. The local polynomial regression lines (LOESS) indicate that hypermutation rates for HBeAg-negative patients are higher, on average, than for HBeAg-positve patients. Positions are with respect to reference strain AM282986. The HBV genomic regions were added to plots A and B.

for HBeAg-negative patients, respectively. Thus, the highest median hypermutation rate for HBeAg-negative patients was 14.6 times higher than that of HBeAg-positive patients.

Genomic regions. Hypermutation was nonuniformly distributed across the genome (Figure 5.2). The highest hypermutation rates were observed between nucleotide positions 600 and 1800 of the HBV genome. In this genomic region the average hypermutation rate was $0.067 \pm 0.2\%$ for HBeAg-positive and $0.35 \pm 0.6\%$ for HBeAg-negative patients, respectively. Intermediate hypermutation rates were observed between nucleotide positions 1 and 600 with an average hypermutation rate of $0.03 \pm 0.01\%$ and $0.1 \pm 0.2\%$ for HBeAg-positive and HBeAg-negative patients, respectively. Between nucleotide positions 1800 and 3221 the observed average hypermutation rates were of $0.01 \pm 0.07\%$ and $0.05 \pm 0.2\%$ for HBeAg-positive and HBeAg-negative patients, respectively. The peak location of high hypermutation rates was independent of the applied combination of hypermutation criteria, e.g. an analysis in which a cutoff value of ten G-to-A mutations was used yielded very similar results.

Dinucleotide context. The dinucleotide context of the G-to-A hypermutations indicated preferred editing in the GA and the GG context and little editing in the GT context for amplicons 1 to 4 (Figure 5.3). For amplicons 5 to 7 the data was very sparse, as the median hypermutation rates were near zero in this genomic region. A3 deaminases prefer to cause mutations in the GG and the GA dinucleotide context. Clonal analysis of the hypermutated sequences revealed that, depending on the genomic region (amplicons 1 to 7), 74% to 89% of the hypermutated sequences mapped to the editing dinucleotide contexts

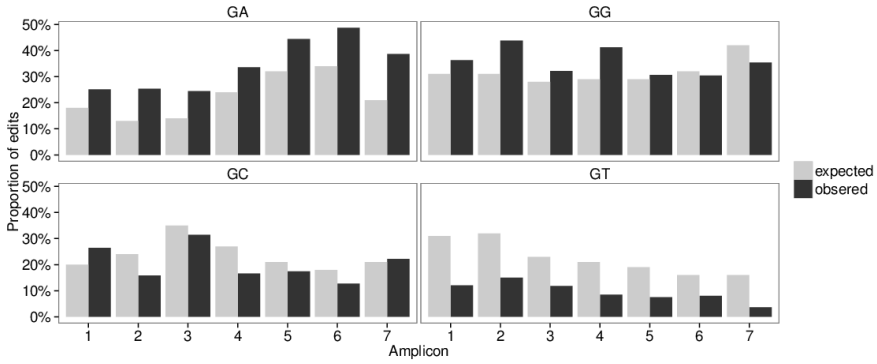

Figure 5.3: Hypermutated sequences were analyzed with respect to the dinucleotide context, in which the G-to-A mutations were found. Hypermutation was defined by at least four G-to-A mutations per read and a G-to-A preference of at least 70%. The actual number of mutations (black) that occurred in the dinucleotide contexts GA, GG, GC, and GT was compared to the relative frequency of the respective dinucleotides in the reference strain AM282986 (grey).

of typical A3 deaminases (Figure 5.4). Thus, after accounting for dinucleotide frequencies in the reference sequence editing was observed more frequently in the GG and the GA context than in the GC and GT context.

Correlation with patient characteristics. Hypermutation rates were correlated with important patient characteristics. The results of the statistical tests are summarized in Table 5.3. Hypermutation rates were computed using sequence data from amplicon 3 (nucleotide positions 1093 to 1660) which overlaps with the peak location of hypermutation. We found that G-to-A hypermutation rates were significantly correlated with the degree of fibrosis for HBeAg-negative patients ($P = 0.004$). This did not hold for HBeAg-positive patients ($P = 0.09$). Figure 5.5 shows increasing hypermutation rates for both HBeAg-positive and HBeAg-negative patients with increasing degree of fibrosis. Additionally, we found that G-to-A hypermutation rates are correlated with the patients age for HBeAg-negative patients ($P = 0.04$). Nonetheless, in a multivariate analysis including degree of fibrosis, age, and hypermutation rates age was no longer significantly associated with hypermutation rates ($P = 0.07$). Other patient characteristics including sex, ethnicity, HBV genotype, ALT, and viral load did not reveal significant associations.

Correlation with escape mutations. Several studies suggested that G-to-A hypermutation might be one of the driving forces of HBeAg seroconversion due to mediation of escape mutations (Noguchi et al., 2009; Turelli et al., 2004; Vartanian et al., 2010). We found that the relative prevalence of the G1764A ($P = 0.0002$) mutation was significantly associated with G-to-A hypermutation rates, which again were computed using amplicon 3 sequences only, in HBeAg-positive patients; no association was found for the mutations G1896A ($P=0.8$) and A1762G ($P = 0.4$). G-to-A hypermutation was also suspected to be relevant for the development of immune escape in the "a" determinant of HBsAg, e.g. mutations sG145R or sG145E (both mediated by G-to-A transitions) (Turelli et al., 2004; Vartanian et al., 2010). We found that for HBeAg-positive patients the relative prevalence of these mutations was strongly correlated with G-to-A hypermutation rates

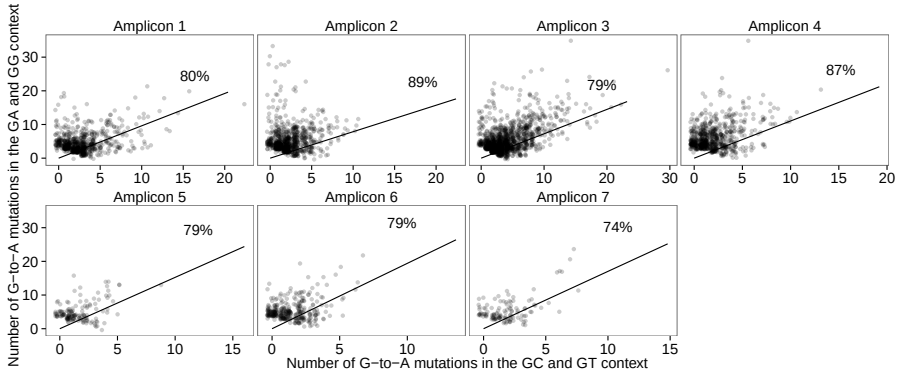

Figure 5.4: Hypermutated sequences were analyzed at the single-read level with respect to the dinucleotide context in which the G-to-A mutations were found. Hypermutation was defined by at least four G-to-A mutations per read and a G-to-A preference of at least 70%. Each dot in the scatterplot refers to one hypermutated sequence in the 454-data. The x-axis denotes the number of G-to-A mutations in the GC and GT dinucleotide context per read and the y-axis displays the number of G-to-A mutations in the GA and GG context. The diagonal lines extending from the origin account for the frequencies of the dinucleotides in the reference strain AM282986 of the respective genomic region (amplicons 1 to 7) and separate the scatterplot into an upper-left region, in which mutations in the GA and GG context are more prevalent than expected, and a lower-right region, in which mutations in the GC and GT context are more prevalent than expected. The percentage numbers give the fraction of sequence reads in the respective genomic regions that express higher mutation rates in the GA and GG context and, thus have likely been edited by A3 deaminases.

	HBeAg-positive		HBeAg-negative	
Patient characteristic	P-value (HMR1)	P-value (HMR2)	P-value (HMR1)	P-value (HMR2)
Age	0.442	0.701	**0.043**	0.161
Sex	0.794	0.844	0.365	0.448
Ethnicity	0.613	0.161	0.495	0.189
HBV genotype	0.787	0.876	1.000	0.709
ALT	0.150	0.093	0.881	0.890
HBV DNA	0.425	0.481	0.974	0.871
Degree of Fibrosis	0.087	0.106	**0.004**	**0.016**

Table 5.3: Hypermutation rates associated with patient characteristics. HMR1: hypermutation defined by at least four G-to-A mutations and G-to-A preference of at least 70% (amplicon 3 reads only). HMR2: hypermutation defined by at least ten G-to-A mutations and G-to-A preference of at least 70% (amplicon 3 reads only).

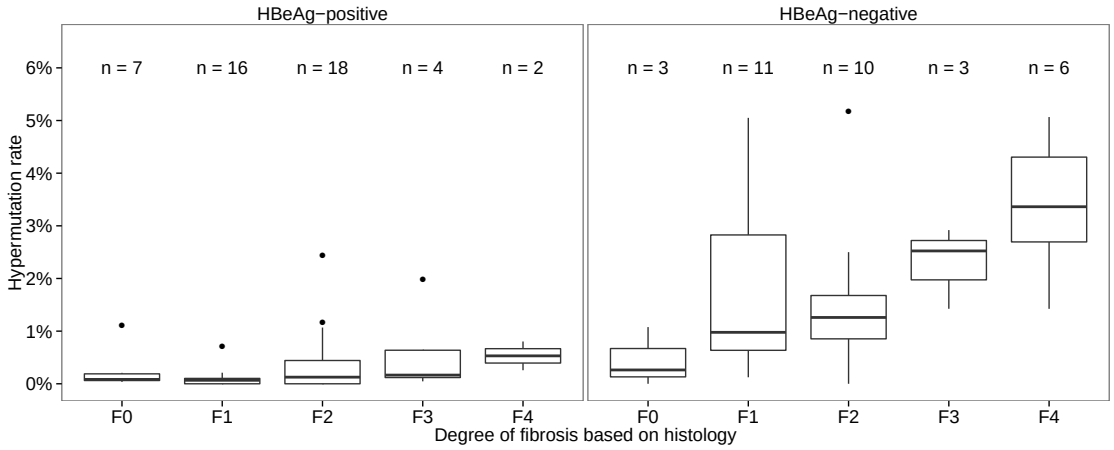

Figure 5.5: Hypermutated reads were defined by at least four G-to-A mutations and a G-to-A preference of at least 70% (amplicon 3 sequences only). Hypermutation rates increase with more severe fibrosis indicating higher activity of A3 deaminases in fibrotic patients.

($P<0.0001$). For HBeAg-negative patients absent or weaker correlations were observed for sG145R ($P=0.9$) or sG145E ($P=0.01$), respectively.

5.4 Discussion

By analyzing second-generation sequencing data of the complete HBV genome of 47 HBeAg-positive and 33 HBeAg-negative treatment-naïve patients, we identified a nonuniform distribution of G-to-A hypermutations across the genome and significantly different hypermutation rates for HBeAg-positive and HBeAg-negative patients. We also found that hypermutation rates were significantly correlated with the degree of fibrosis in HBeAg-negative hepatitis.

Dinucleotide context. Hypermutated sequence reads were analyzed at the clonal level with respect to the dinucleotide context of the observed G-to-A exchanges. Depending on the genomic region we observed that 74% to 89% of the hypermutated reads contain more GG-to-AG and GA-to-AA mutations compared to GC-to-AC and GT-to-AT exchanges than one would expect based on the dinucleotide frequencies of the HBV genome. This indicates that the majority of hypermutations is due to the action of A3 deaminases which display the respective editing profile.

Distribution across the genome. A peak location of hypermutated reads was observed between nucleotide positions 600 and 1800 (see Figure 5.2), which does not result from uneven coverage during sequencing but correlates with the replication cycle of HBV as proteins of the APOBEC family prefer to deaminate single-stranded DNA. In fact, mature HBV particles contain a genome in which nucleotide positions between approximately 138 and 947 are partially single-stranded and nucleotide positions from 947 to the (+)-DNA strand synthesis primer site are almost exclusively single-stranded. The predominant (+)-DNA strand primer site of HBV is the direct repeat element DR2 (nucleotide positions

1824 to 1834). Thus, the detected hypermutation rates (Figure 5.2) are consistent with the prevalence of single-stranded DNA *in vivo*: high hypermutation rates were found where HBV DNA is almost exclusively single-stranded (between nucleotide positions 947 and DR2) and intermediate hypermutation rates were observed where HBV DNA is partially single-stranded (between nucleotide positions 138 and 947). Hypermutated genomes with low frequency were also located between nucleotide positions 1800 to 3221. This might hint at APOBEC activity on the HBV genome before the synthesis of the (+)-DNA strand, thus before envelopment. This situation is in contrast to e.g. editing the genome of the human immunodeficiency virus, which can only be targeted after cell entry and reverse transcription of the viral RNA.

HBeAg status. A key finding of our study is that the median G-to-A hypermutation rates of HBeAg-negative patients were more than 10-fold higher than those of HBeAg-positive patients. This finding is likely due to the fact that HBeAg-negative patients have undergone HBeAg seroconversion implying previous mature immune system activity. Alternatively, suppression of the immune modulator HBeAg may have resulted in increased immune pressure. Our finding contradicts a study based on cloning and conventional sequencing of ten chronically infected patients in which hypermutated genomes were only found in HBeAg-positive but not in HBeAg-negative patients (Noguchi et al., 2005). Increasing hypermutation rates in HBeAg-positive chronic hepatitis were previously associated with HBeAg loss (Noguchi et al., 2009). Interestingly, we found a significant correlation of the hypermutation rates with the relative prevalence of the G1764A mutant but not with the G1896A precore stop codon mutation. The G1764A mutation, which is known to be an early indicator of HBeAg loss, is located within the genomic region of high hypermutation rates but the G1896A precore stop codon mutation is not (Chu et al., 2003; Hussain et al., 2003). Thus, our data substantiate the hypothesis that the activity of A3 deaminases supports the emergence of HBeAg escape mutations and might drive HBeAg seroconversion.

Fibrosis. Several cytosine deaminases including A3B, A3C, A3G, A3H, and AID are upregulated in HBV associated cirrhotic liver tissue (Vartanian et al., 2010). Our data indicate that upregulation of these genes results in a significantly increased number of mutated genomes in sera of HBeAg-negative patients. This association was not significant for HBeAg-positive patients but a similar trend was observed (Figure 5.5).

Conclusions. Our data provide new insights into the G-to-A hypermutation patterns of the HBV genome. Hypermutation significantly depends on the genomic region, the patients' HBeAg and fibrosis status. The observation that hypermutation rates are 10-fold higher in HBeAg-negative patients and that the relative prevalence of the core promoter mutation G1764A correlates with hypermutation rates for HBeAg-positive patients indicates an important association of hypermutation mediated by A3 deaminases with the natural progression of chronic hepatitis B infections both in term of HBeAg seroconversion and disease progression towards cirrhosis.

6 Predicting Treatment Response to Interferon

In science it often happens that scientists say, "You know that's a really good argument; my position is mistaken," and then they would actually change their minds and you never hear that old view from them again. They really do it. It doesn't happen as often as it should, because scientists are human and change is sometimes painful. But it happens every day. I cannot recall the last time something like that happened in politics or religion.

(Carl Sagan, 1987)

Interferons facilitate finite treatment durations of only 48 weeks, while treatment duration with nucleos(t)ide analogues is unpredictable and nucleos(t)ide analogues often require life-long administration. This benefit of interferon is accompanied by two major downsides. First, almost all patients treated with interferon experience severe adverse effects, for instance fatigue, anorexia, emotional lability, hair loss, and exacerbation of autoimmune illnesses. Second, sustained response is only achieved in a fraction of patients. Rates of sustained response vary between 19% and 30% depending on the patients' HBeAg status and the definition of sustained response. Thus, predicting the probability of response before start of treatment is of high clinical relevance. We review the state of the art of predicting treatment response to interferon in Section 2.3.1. In short, high serum alanine aminotransferase levels, low HBV DNA levels, the presence of HBV genotypes A or B, and progressed liver inflammation are pretreatment indicators of sustained response but provide only limited evidence. On-treatment indicators have been shown to be more informative. Patients that do not exhibit a significant decline of HBsAg and HBV DNA levels after twelve weeks of treatment have very low probability of achieving sustained response. Thus, interferon therapy should be stopped for these patients as the treatment will likely not be successful.

In Chapter 5 we analyzed G-to-A hypermutation patterns using full-genome second-generation sequencing data. We saw that G-to-A hypermutation can be associated with the natural course of hepatitis B. In this Chapter we use the exact same data obtained from 47 HBeAg-positive and 33 HBeAg-negative carriers to investigate whether or not pretreatment G-to-A hypermutation patterns are informative of sustained response to interferon. We hypothesized that G-to-A hypermutation might serve as a surrogate marker for assessing the activity of the innate immune system, which may further be stimulated by treatment with interferon. We found that only the prevalence of hypermutated genomes with a GG-to-AG preference correlated highly with treatment response for HBeAg-negative carriers. This finding could not successfully be verified on a validation cohort. Thus, G-to-A hypermutation rates are not predictive of treatment response.

This Chapter is structured as follows. In Section 6.1 we present the methodical details of our study. In particular, we explain the approach used to derive features based on

G-to-A hypermutated genomes and the supervised learning algorithm we employed. In Section 6.2 we discuss the results of our analysis (referred to as the original study) and derive a prediction model for HBeAg-negative patients. The model was not predictive on a validation cohort (referred to as the validation study), as detailed in Section 6.3. We conclude this Chapter with a discussion in Section 6.4. The work presented in this Chapter was joint work with Prof. Andreas Erhardt (formerly Heinrich-Heine-University, Düsseldorf, currently Petrus Hospital, Wuppertal) who designed the study and provided patient data of the original study, and Prof. Carsten Münk (Heinrich-Heine-University, Düsseldorf) who supported and guided the analysis. Patient data and material for the validation study was supplied by André Boonstra and Harry L. A. Janssen (both Erasmus MC-University Medical Center, Rotterdam).

6.1 Material and Methods

Original study. Second-generation sequencing (Roche/454 pyrosequencing) was performed on pretreatment serum samples from 47 HBeAg-positive and 33 HBeAg-negative patients treated with pegylated interferon α-2a (Lau et al., 2005; Marcellin et al., 2004). Duration of therapy was 48 weeks. After 24 weeks of follow-up, sustained response was evaluated according to the study protocol: HBV DNA level below 20,000 copies per milliliter, normalization of ALT (below 30 international units per milliliter), and in case of HBeAg-positive patients HBeAg seroconversion. Patient characteristics are detailed in Table 5.1.

Validation study. To verify the prediction model that was developed based on the results of the original study a validation study was performed. 45 treatment-naïve HBeAg-negative patients were enrolled in the validation study. Patients were treated with pegylated interferon α-2a at the Erasmus MC-University Medical Center in Rotterdam. Duration of therapy and definition of sustained response was in accordance with to the original study.

Sequencing data. Preparation of sequencing data for the original study is discussed in Section 5.2. Sequencing for the validation study was performed using Illumina/MiSeq sequencing. Two amplicons per patient sample covering NT positions 57 to 1824 were prepared as in Zhang et al. (2007). Preparation of sequencing libraries involved equimolar pooling of amplicons, which were fragmented and tagged using a Nextera DNA sample preparation and index kit from Illumina. The resulting sequencing libraries were quantified on a 2100 Bioanalyzer (Agilent Technologies) and diluted to 10 picomole per liter for amplification and sequencing on an Illumina/MiSeq sequencer using the 2 times 250 bases paired-end sequencing protocol. A cutoff value of 25 for the phred-equivalent quality scores was used to clip parts of the raw sequence data of inferior quality. Otherwise pre-processing of sequencing data was performed in the same manner as for the 454-data (Section 5.2).

Hypermutation grid. Each sequencing read was classified as normal or hypermutated based on the G-to-A preference (proportion of G-to-A mutations divided by the total number of mutations) and the number of G-to-A mutations with individual cutoffs. Different cutoff values 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90% for the G-to-A preference and 2, 3, 4, 5, 6, 7, 8, 9, and 10 for the minimum number of G-to-A mutations were tested. Each preference cutoff was combined with each number of mutations cutoff, resulting in 81 definitions of hypermutation. These combinations of criteria were

evaluated on multiple window sizes along the reads (25, 50, 75, 100, 200, 300 bases, and full read length), resulting in 567 definitions of HMRs. A read was classified as hypermutated if there was at least one window of the respective size along the read for which both hypermutation criteria were fulfilled.

Hypermutation context. Proteins of the APOBEC family display different dinucleotide preferences for cytosine deamination (Section 5.1.3). Therefore, extensions of the G-to-A preference statistic were implemented to account for the dinucleotide context. For each nucleotide X (A, C, G, or T) the GX-to-AX preference was computed and combined with the cutoff values discussed before.

Feature selection. An univariate scoring scheme based on subsampling identified the hypermutation rates that best correlated with treatment response (Guyon and Elisseeff, 2003). 50% of the training data were sampled 200 times without replacement. For each iteration the association of each HMR with the response was measured in terms of the area under the receiver operating characteristic curve (AUC). The final ranking was performed using the frequency of the HMRs in the top 5% of the AUC-ranked list. The four top-scoring hypermutation rates were used in the subsequent prediction. The exact number of selected HMRs was not critical. Almost identical results were obtained with three to nine HMRs.

Clinical features. In addition to the top-ranked hypermutation rates other clinical parameters, such as pretreatment HBV DNA levels and pretreatment ALT levles, were used as relevant indicators of treatment response.

Performance assessment. The performance of the statistical preciction procedure was evaluated using 20 times ten-fold cross validation. Feature selection, as described above, was performed within each fold using only the current training data. All features were normalized to have zero mean and standard deviation one. A linear support vector machine model (as introduced in Section 2.5.3) was trained using the selected features to predict the test data of the current fold. Class weights were computed within each fold and passed to the SVM to account for unbalanced class distributions. Nested cross validation was implemented to regularize model complexity. For each patient the probability of treatment success was predicted. Finally, the AUC was computed and used as primary quality criterion. Treatment predictions were made by computing the maximum distance of all nonresponders (of the current training data) to the separating hyperplane. The current test data was classified using this decision boundary. These predictions were used to compute the prediction accuracy.

6.2 Original Study

Baseline characteristics. Sustained response to pegylated interferon α-2a in the HBeAg-positive and HBeAg-negative cohort was achieved by six of 47 patients and by eight of 33 patients, respectively. The responders in the HBeAg-positive cohort showed non-significantly lower pretreatment HBV DNA levels ($P = 0.80$, Wilcoxon rank-sum test) and significantly higher pretreatment ALT levels ($P = 0.04$, Wilcoxon rank-sum test) compared to the nonresponders. In the HBeAg-negative cohort pretreatment HBV DNA levels were significantly lower for the responders ($P = 0.05$, Wilcoxon rank-sum test), while pretreatment ALT levels were non-significantly higher ($P = 0.98$, Wilcoxon rank-sum test). The

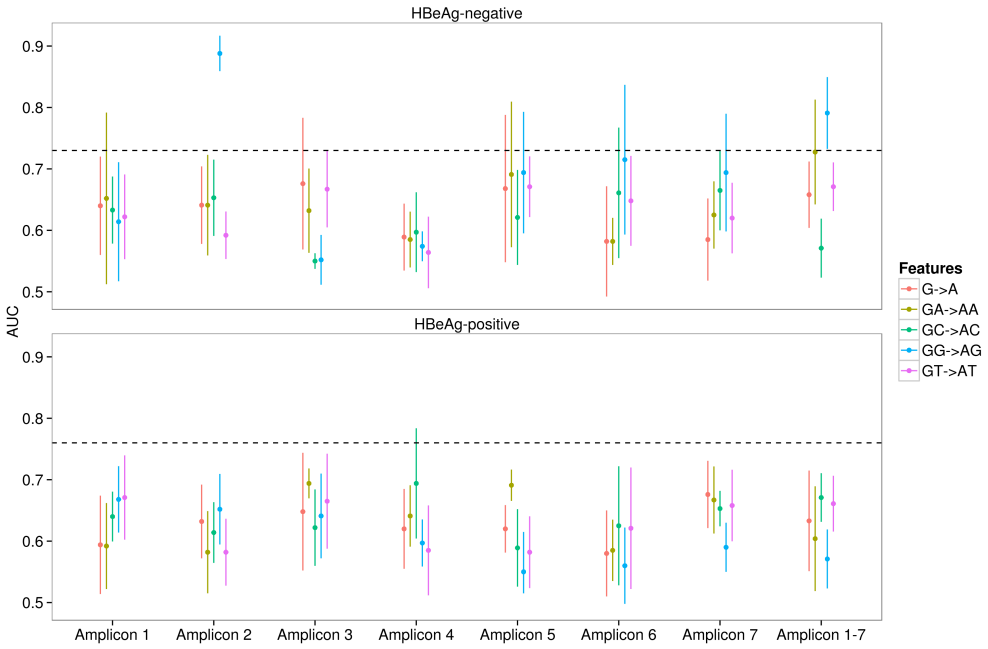

Figure 6.1: Prediction results for HBeAg-negative and HBeAg-positive cohort with respect to the genomic region and the GX-to-AX hypermutation features. The dashed lines indicate the performances of the baseline methods in terms of AUC.

dependency of HBV DNA on the sustained response to pegylated interferon α-2a for the HBeAg-negative patients in terms of the AUC was 0.73 and for HBeAg-positive patients the dependency of ALT with the sustained response in terms of the AUC was 0.76. These associations represent baseline models for the prediction of treatment response. The prediction performance may further be improved using features extracted from second-generation sequencing data based on the hypermutation patterns.

Prediction of sustained response. To predict sustained response, we computed various hypermutation rates using a combination of criteria: the number of G-to-A mutations and the GX-to-AX preference evaluated on various window sizes along each read. We then used the subsampling-based feature selection algorithm to identify relevant definitions of hypermutation rates and calculated SVM models to predict sustained response. The analysis was performed with eight different subsets of the sequence data. We used sequence data that was generated using each amplicon (1 to 7, see Table 5.2) separately and, in addition, the set of all available sequence data (full genome). Additionally, separate models were calculated for the five (di)nucleotide motifs (G-to-A, GA-to-AA, GC-to-AC, GG-to-AG, and GT-to-AT). This gave rise to eight times five equals 40 configurations for both the HBeAg-positive and the HBeAg-negative cohorts. Each configuration used 567 hypermutations rates based on the combination of hypermutation criteria as described in the Methods section. Figure 6.2 summarizes the prediction performances for both patient cohorts with each of the 40 configurations in terms of the AUCs.

HBeAg-positive cohort. The features based on hypermutation rates did not facilitate the prediction of treatment response to interferon in the HBeAg-positive cohort.

None of the 40 configurations with respect to the feature subset and the genomic region exceeded the prediction accuracy of the baseline method (correlation of the pretreatment ALT level with the response (AUC = 0.76)).

HBeAg-negative cohort. The ability to predict the response to interferon in the HBeAg-negative cohort strongly depended on both the genomic region and the features in use. Best performance was achieved using GG-to-AG HMRs based on sequence data from amplicon 2 (nucleotide positions 524 to 1197). The AUC of this model was 0.89±0.03. Compared to the baseline model (correlation of the pretreatment HBV DNA level with the response (AUC = 0.73)), only one other configuration showed minor improvements in prediction accuracy, namely, GG-to-AG HMRs based on the full genome with an AUC of 0.79±0.06. The mean prediction performance of the other configurations did not exceed the AUC of the baseline method. The hypermutation rates based on G-to-A, GA-to-AA, GC-to-AC, and GT-to-AT preferences do not significantly contribute to prediction accuracy.

Informative predictors for the HBeAg-negative cohort. We analyzed in more detail the cutoff values that were inferred by the model using amplicon 2 sequences (NT 524 to 1197) and GG-to-AG hypermutation rates. For this purpose, we examined the coefficients of the respective SVM model. The most relevant feature was the GG-to-AG HMR defined on the basis of at least four G-to-A mutations and a preference cutoff of 50% on a window size of 300 followed by the pretreatment HBV DNA levels. Figure 6.2A illustrates the support vector machine model using these two features. Linear separation of responders and nonresponders with only one misclassification can be achieved using HBV DNA levels and the GG-to-AG HMR.

We computed the distribution of GG-to-AG hypermutation rates for the genomic regions obtained from different amplicons separately for both responders and nonresponders. For amplicons 1, 4, 5, 6, and 7 the median GG-to-AG hypermutation rates were zero. This is in accordance with the very low general G-to-A hypermutation rates of these regions. For amplicon 2, responders and nonresponders showed different GG-to-AG hypermutation rates with median rates of 0.19% and 0% (P = 0.0002, Wilcoxon rank-sum test), respectively. For amplicon 3 the GG-to-AG hypermutation rates of the responders ranged from 0% to 0.36% (median 0.12%) and the GG-to-AG hypermutation rates of the nonresponders ranged from 0.00% to 0.40% (median 0.07%). Thus, the hypermutation rates in the genomic region covered by amplicon 3 did not allow for separation of the two patient groups (P = 0.59, Wilcoxon rank-sum test). We conclude that only GG-to-AG HMRs found in the genomic area of NT positions 524 to 1197 were informative of sustained response.

Prediction model for the HBeAg-negative cohort. We found that the performance of the prediction model could be slightly improved by removing the pretreatment ALT from the set of features. The AUC of the model trained on amplicon 2 and GG-to-AG HMRs using only the pretreatment HBV DNA as secondary criterion was 0.90±0.03. Therefore, our final prediction model only included the pretreatment HBV DNA levels and the four top-ranked GG-to-AG hypermutation features. To predict sustained response on new data the weighted sum of the four HMRs and the pretreatment HBV DNA using the definitions and weights in Table 6.1 need to be computed. Positive response predictions must have a weighted sum that exceeds a cutoff value of -0.14053. On the HBeAg-negative cohort this model's accuracy amounted to 73±3%.

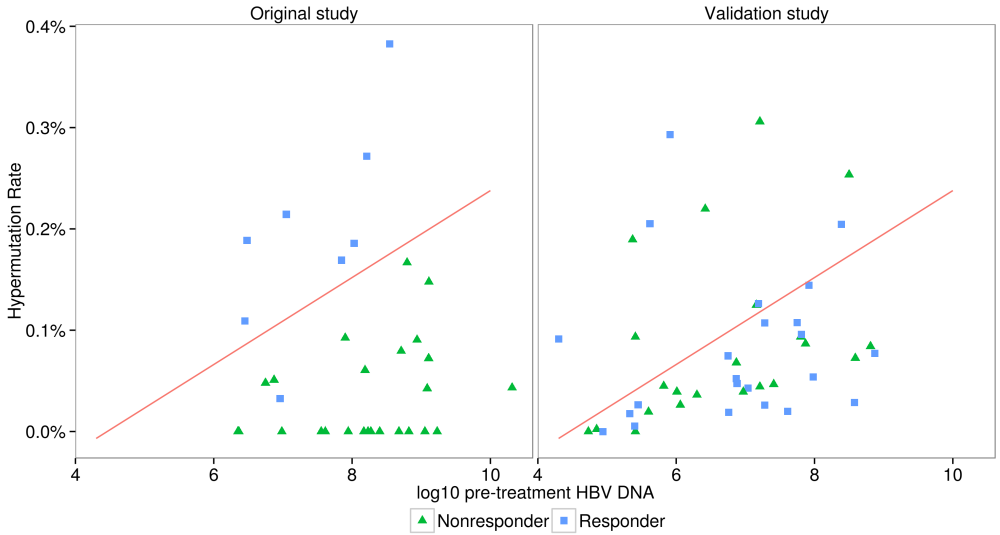

Figure 6.2: Separation of responders and nonresponders based on the GG-to-AG hypermutation rates and the HBV DNA levels in the original study (A) and the validation study (B). (A) In the original study very good separation with only one misclassification can be observed. (B) The GG-to-AG hypermutation rates and the HBV DNA levels did not allow for separation of responders and nonresponders in the validation cohort.

Predictor	Predictor definition	Predictor weight
GG-to-AG HMR1	Primer pair 2; window size 300; G-to-A mutations at least 4; GG-to-AG preference at least 50%	2.76028
GG-to-AG HMR2	Primer pair 2; window size 100; G-to-A mutations at least 4; GG-to-AG preference at least 70%	2.74560
GG-to-AG HMR3	Primer pair 2; window size 75; G-to-A mutations at least 4; GG-to-AG preference at least 70%	3.34946
GG-to-AG HMR4	Primer pair 2; window size 50; G-to-A mutations at least 4; GG-to-AG preference at least 50%	3.60747
HBV DNA	Log10 copies per milliliter	-0.02012

Table 6.1: Prediction model specification derived from the original study. Positive response predictions had a weighted sum that exceeded the cutoff value of -0.14053.

6.3 Validation Study

To test the prediction model derived on a rather small sample size of 33 patients with only eight responders and 25 nonresponders a validation study was performed. Therefore, second-generation sequencing data covering NT positions 524 to 1197 were generated for a validation cohort. In total 45 HBeAg-negative patients treated with pegylated interferon α-2a were enrolled in the validation study, 23 of which were responders according to the definition of sustained response.

Sequencing data were processed according to the original study. Due to the low prevalence of the GG-to-AG hypermutated genomes it was essential to provide sufficient coverage for the genomic region of interest. The median coverage (number of reads per patient sample and nucleotide position) in the validation cohort had a mean of 60037 ± 17869. Thus, the coverage was approximately 41 times higher compared to the original study.

The accuracy of the prediction model as specified in Table 6.1 on the validation data set amounted to 46.7%. The correlation of the distance to the SVM hyperplane with the actual response in terms of AUC was 0.56. Thus, the model performs no better than random guessing on the validation data set and does not facilitate the prediction of treatment response.

6.4 Discussion

By analyzing pretreatment second-generation sequencing data of the complete HBV genome of 47 HBeAg-positive and 33 HBeAg-negative patients, we investigated whether G-to-A hypermutation patterns might serve as a novel predictor for sustained response to pegylated interferon α-2a. Sustained response was defined as HBV DNA level below 20,000 copies per milliliter combined with normalization of ALT (below 30 international units per milliliter) after 24 weeks of follow-up. For HBeAg-positive patients HBeAg seroconversion was also required for sustained response. Initial promising results for HBeAg-negative patients using GG-to-AG hypermutated sequences could not be successfully verified on a validation cohort.

Various definitions based on the combination of hypermutation criteria were employed to derive a feature set for predicting treatment response to interferon. Prediction performance was evaluated using 20 times ten-fold cross validation and expressed in terms of the AUC. Elevated pretreatment ALT levels and low HBV DNA levels are established indicators of sustained response for HBeAg-positive and HBeAg-negative patients, respectively.

HBeAg-positive cohort. Using hypermutation rates based on second-generation sequencing data did not improve treatment response prediction performance compared to the AUC of the baseline method (pretreatment ALT levels) for HBeAg-positive patients. This negative result might be the consequence of three major limitations of our study design. First, the low sample size of only 47 HBeAg-positive patients might undermine the identification of relevant features. Second, positive responders were underrepresented in our patient cohort as only six of 47 achieved sustained response. Unbalanced distributions of labels have been shown shown to hamper optimal classifier performance (Japkowicz and Stephen, 2002). We accounted for these two limitations by the use of SVMs, which have been shown shown to perform well with low sample sizes and unbalanced class distribu-

tions, by using class weights to assign higher weights to the underrepresented class, and with a rigorous univariate feature selection based on resampling. Third, hypermutation rates for HBeAg-positive patients are very low in relation to the coverage. As presented in Chapter 5, median hypermutation rates for segments of 100 base-pair along the genome did not exceed 0.047% while the median coverage was 1,480. Thus, for half of the samples only $1,480 \cdot 0.047/100 \approx 0.7$ can be expected within each segment.

HBeAg-negative cohort. Using the original data set of 33 HBeAg-negative patients we found that GG-to-AG hypermutations, which occur at frequencies of approximately 10^{-3}, in combination with HBV DNA levels allowed accurate prediction of treatment response to interferon with an AUC of 0.9 ± 0.03. In cell culture, a high number of GG-to-AG mutations is characteristic for the antiviral activity of A3G on HBV. Other A3 deaminases have a weaker or different dinucleotide context associated with editing and, e.g. A3B, A3C, and A3F induce G-to-A hypermutations also in other dinucleotide contexts. This might explain the high prevalence and low association with treatment response of hypermutated reads that do not exhibit a strong GG-to-AG preference. However, the prevalence of GG-to-AG hypermutated genomes suggests expression and antiviral activity of A3G on HBV prior to treatment with interferon. The detected difference in the GG-to-AG HMR between responders and nonresponders might be a marker for inter-patients variability of expression of A3G or, more generally, of interferon stimulated genes that inhibit HBV.

GG-to-AG HMRs were observed at very low frequencies ($\approx 10^{-3}$). The number of GG-to-AG hypermutated reads ranged between one and eleven (median of 2.5) for responders and between zero and four (median of zero) for nonresponders. Hence, providing sufficient sequencing depth was critical for the validation study. Therefore, Illumina/MiSeq sequencing was applied which provides an approximately 60-fold lower price per megabase (Loman et al., 2012).

Validation study. The prediction model based on pretreatment GG-to-AG hypermutation rates and HBV DNA levels was not successfully verified on the validation cohort of 45 patients. Prediction accuracy amounted to 46.7% and the AUC was only 0.56. Thus, model performance was no better than random.

The change of the sequencing technology, which subsequently also implied a different set of primers, might have introduced several kinds of biases that might have altered the result of the analysis. Even though this cannot be excluded, it seems unlikely as both sequencing technologies (Roche/454 pyrosequencing and Illumina/MiSeq) have been proven to provide state-of-the-art sequencing results. Read length is critical to identifying hypermutated reads as the criteria of hypermutation are applied at (sub) read level. Read length of 400 bases as provided by Roche/454 pyrosequencing is not required. Our prediction model gave high weights to GG-to-AG HMRs evaluated on window sizes of 50, 75, and 100 bases (Table 6.1). The Illumina/MiSeq sequencing platform produces reads of 250 bases length, which is sufficient to compute GG-to-AG hypermutation rates. Consequently, the relative frequency of GG-to-AG hypermutation rates did not differ notably between the original and validation data sets (see Figure 6.2). Nevertheless, GG-to-AG hypermutation rates did not allow separation of responders and nonresponders in the validation cohort.

Thus, we think in the original study the GG-to-AG HMRs correlated with treatment response just by chance due to the low sample size of only 33 patients with only eight

responders. Studies performed on small sample sizes always bear the risk of producing false positive results (Hackshaw, 2008). We maintained a consistent separation of training and test data when selecting the features and evaluating the performance on the original data set. The validation of our prediction method showed that the GG-to-AG hypermutation features did not generalize to the patient samples of the validation study. To further investigate this, the prediction procedure (including feature selection and SVM prediction in a ten-fold cross-validation setting) was repeated 500 times with randomized response labels on the original data set. In three (0.6%) repetitions a performance in terms of AUC that exceeded 0.89 was observed.

Conclusions. Our hypothesis that G-to-A hypermutation patterns might serve as a novel indicator to predict sustained response to interferon could not be successfully validated on our patient cohorts of 47 HBeAg-positive and 33 HBeAg-negative patients. The study suffered from several major limitations, low sample size and underrepresentation of positive responders. Additionally, the low prevalence of G-to-A hypermutated genomes among HBeAg-positive carriers in relation to the coverage also impaired the evaluation of our hypothesis. Despite these limitations, our analysis provides evidence that G-to-A HMRs do not facilitate the prediction of interferon treatment response.

7 Conclusions

Real Biology has at least 50 more interesting years.

(James Watson, 1988)

We have presented methods for determining and utilizing the quasispecies of the hepatitis B virus in the context of clinical applications. Determining the viral quasispecies relies on data obtained from DNA sequencing technologies of the first or second generation, which were analyzed using methods from the field of statistical learning and probabilistic reasoning. This Chapter provides summarizing remarks and offers possible directions to extend the presented work.

7.1 Summarizing Remarks

The main objective of our work was to develop statistical methods to analyze the quasispecies of the hepatitis B virus to improve the clinical care of infected patients.

In particular, we developed the first *in silico* genotyping method that can identify and genotype HBV intra- and intergenotype dual infections. The method showed high accuracy on synthetic test data. Intergenotype dual infections were correctly genotyped in 98.5% to 100% of the test cases depending on the genomic region, and intragenotype dual infections were identified correctly in 44.2% to 48.6% of the test samples. Other state-of-the-art genotyping methods do not account for dual infections and might underestimate the risk of interferon therapy failure if one of multiple genotypes is not recognized. The method was used to assess the frequency of dual infections in routine diagnostics. Eight (3.3%) intergenotype and four (1.7%) intragenotype dual infections were identified in our patient cohort (n = 241). Clonal sequence analysis performed on three patient samples confirmed our predictions and revealed very complex compositions of the quasispecies, which involved multiple distinct recombinant forms in each patient. The dual infection model was integrated into the web-service geno2pheno$_{[\mathrm{HBV}]}$ and is freely available to be used with HBV sequences generated within the clinical routine.

We presented the first approach that utilized peak height profiles from Sanger sequencing chromatograms to infer linkage information about nearby ambiguous positions. A generative model was employed to compute the expected peak heights of a mixture by combining the chromatograms of the underlying haplotypes. The model was shown to infer the haplotypes present in mixtures with high accuracies of 97.4% and 84.5% on two *in vitro* test-sets. The effectiveness of our method shows that short-range linkage information can be inferred from sequencing chromatograms with no further assumptions on the mixture composition. Our model provides new insights into the established and widely used Sanger sequencing technology and facilitates the estimation of the relative frequencies of nearby ambiguous sequence positions, thus it overcomes the limitations of previous

methods. Nevertheless, the model relies on the availability of the chromatograms of all possible haplotypes in the mixture, which limits its widespread applicability. Additionally, we showed that the approach has several significant limitations. First, it is not possible to reliably detect minor variants with a relative frequency of no more than 10%. Second, the model cannot distinguish between mixtures of two or four clonal variants if one of two sets of linear constraints is fulfilled. Third, linkage can only be inferred over a relatively short range of maybe up to five bases. This is due to the locality of the effect of sequence context-dependent incorporation of dideoxynucleotides. Prediction accuracy was already impaired on a test set, in which the positions under analysis were separated by three bases.

The advent of second-generation sequencing technologies offers new perspectives on the viral quasispecies of HBV. We conducted the first full-genome analysis of the G-to-A hypermutation patterns of the HBV genome. We studied hypermutation, mediated by APOBEC deaminases, with respect to clinical patient characteristics, which has not been done before using second-generation sequencing data. We revealed associations of hypermutation with the replication cycle of HBV and with the natural progression of chronic hepatitis B in terms of HBeAg seroconversion and disease progression towards cirrhosis. We showed that hypermutation significantly depends on the genomic region as well as the patients' HBeAg and fibrosis status. To detail this, we provided actual numbers of the relative prevalences of hypermutated genomes with respect to these factors, which might serve as reference for further studies.

One of our key observations was that hypermutation rates for HBeAg-negative patients are about ten-fold higher than for HBeAg-positive patients. Additionally, we could demonstrate that in HBeAg-positive patients the relative prevalence of the core promoter mutation G1764A, which is an early indicator of HBeAg seroconversion correlates with hypermutation rates.

Last, we investigated the potential of G-to-A hypermutation patterns to serve as informative pretreatment indicators to predict treatment response to interferon. Various features based on the dinucleotide context of editing and the number of G-to-A mutations were derived. A supervised learning algorithm was applied to compute prediction models for different genomic regions and different subsets of the feature set for HBeAg-positive ($n = 47$) and HBeAg-negative ($n = 33$) patients. The prediction model for the HBeAg-positive cohort did not facilitate the prediction of treatment response, but for HBeAg-negative patients we found that prevalence of GG-to-AG hypermutations in a specific genomic region correlated highly with treatment response on our original data set with an AUC of 0.90 ± 0.03. However, our finding was not successfully validated on a second patient cohort of 45 HBeAg-negative patients. The accuracy of the prediction model based on pretreatment GG-to-AG hypermutation rates amounted to only 46.7% with an AUC of 0.56. We kept a consistent separation of training and test data when selecting the features and evaluating the performance on the original data set. Nevertheless, the validation of our prediction method on the validation cohort showed that the GG-to-AG hypermutation rates did not facilitate the prediction of treatment response to interferon.

7.2 Outlook

One natural direction of further research would be to apply the developed methods to other viruses. The dual infection model was successfully evaluated using synthetic test data to genotype HCV dual infections and HIV dual infections in a Bachelor's thesis (Savenko, 2013). These models have to be validated on sequence data derived from patient sera and could be integrated into freely available web-services.

The dual infection model was developed to genotype dual infections based on first-generation sequencing data. Second-generation sequencing data can represent viral populations with high sequencing depth and provides linkage information over distances up to the read length. At first glance, this facilitates a more direct and more sensitive way of identifying and genotyping dual infections simply by analyzing each read individually. At second glance, several difficulties with this approach become evident involving short reads and conserved regions that might not facilitate reliable genotyping, multiple recombination events that hamper genotyping on single-read level, heterogeneous coverage, and contaminations of various kinds that need to be identified as such to prevent false positive dual infection predictions. Thus, genotyping each read individually seems to shift the burden of deciding whether or not a dual infection is present back to humans to interpret thousands of single-read genotyping results. Another alternative to identifying and genotyping dual infections based on second-generation sequencing data is to make use of quasispecies reconstruction models (reviewed in Beerenwinkel et al. (2012)) to infer the set of underlying haplotypes and perform genotyping on these. This alternative has its own set of difficulties, namely high computational costs implying long run times and error-proneness in case of low prevalences of minority strains or heterogeneous coverage. The basic idea of the dual infections model to use genotype-specific profiles (derived from genotype annotated sequence data) might be of assistance in overcoming these difficulties.

Our method for quantifying the relative frequencies of ambiguous sequence positions and for inferring short-range linkage from Sanger sequencing chromatograms is also not limited to HBV. It can be applied to other viruses or other genomic mixtures for which short-range linkage is of interest. Second-generation sequencing technologies, which are far more sensitive and accurate in determining the composition of mixtures than first-generation methods, superseded our approach in many possible applications. Nevertheless, our methodology is a core building block for resolving the viral quasispecies based on first-generation sequencing data. Our method might be used in conjunction with other sources of information, e.g. patterns of co-occurring mutations derived from sequence databases. To approach this, several difficulties have to be addressed. First, the statistical model and the respective inference procedure need to be extended to include more than two ambiguous positions. Second, the model needs to be applicable in situations in which not all peak height profiles of all possible underlying haplotypes are known, as the number of all possible haplotypes grows exponentially with the number of ambiguous positions under consideration. Third, multiple sources of information and their associated levels of uncertainty have to be incorporated into a joint probabilistic model. Last, the resulting model needs to be computationally tractable.

Finally, we discovered interesting associations between G-to-A hypermutation patterns and the natural progression of hepatitis B. We observed that patients in the HBeAg-

negative chronic phase show about ten-fold higher hypermutation rates than patients in the immune-active phase and that hypermutation might serve as an early indicator of HBeAg seroconversion. Nevertheless, our data could not elucidate during which phases of the chronic infections hypermutations are accumulated and how hypermutation might shape the development of the disease. The analysis of longitudinal second-generation sequencing data of chronically infected HBV carriers could provide additional insights into the development of the quasispecies during the natural progression of hepatitis B.

List of Figures

List of Tables

Acronyms

2ndGS second generation sequencing

3TC Lamivudine

A3 APOBEC3

A3A APOBEC3A

A3B APOBEC3B

A3C APOBEC3C

A3DE APOBEC3DE

A3F APOBEC3F

A3G APOBEC3G

A3H APOBEC3H

AASLD American Association for the Study of Liver Disease

ADV Adefovir

ALT alanine aminotransferase

anti-HBc antibody to the Hepatitis B core antigen

anti-HBe antibody to the Hepatitis e antigen

anti-HBs antibody to the Hepatitis B surface antigen

APASL Asian Pacific Association for the Study of Liver

APOBEC apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like

ASHV artic squirrels hepatitis B virus

ATP adenosine triphosphate

AUC area under the receiver operating characteristic curve

cccDNA covalently closed circular DNA

CHBV crane hepatitis B virus

ddNTP dideoxynucleotide triphosphates

DHBV duck hepatitis B virus

dNTP deoxynucleotide triphosphates

DR direct repeat element

EASL European Association for the Study of Liver

ELISA enzyme-linked immunosorbent assay

ER endoplasmic reticulum

ETV Entecavir

GSHV ground squirrels hepatitis virus

HBcAg hepatitis B core antigen

HBeAg hepatitis B e antigen

HBsAg hepatitis B surface antigen

HBV hepatitis B virus

HBxAg hepatitis B virus X antigen

HCC hepatocellular carcinoma

HCV hepatitis C virus

HHBV heron hepatitis B virus

HHBV heron hepatitis B virus

HIV human immunodeficiency virus

HMM hidden Markov model

HMR hypermutation rate

HPV human papillomavirus

HSV-1 herpes simplex virus type 1

IFN interferon

IU international units

IUPAC International Union of Pure and Applied Chemistry

LdT Telbivudine

LHBsAg large hepatitis B surface antigen

LOESS local polynomial regression lines

MAP maximum a posteriori probability

MHBsAg middle hepatitis B surface antigen

NA nucleos(t)ide analogue

NCBI National Center for Biotechnology Information

NT nucleotide

PAG polyacrylamide gel

PCR polymerase chain reaction

pgRNA pregenomic RNA

PGSND position- and genotype-specific nucleotide distribution

PHBV parrot hepatitis virus

rcDNA relaxed circular DNA

RGHV rose goose hepatitis B virus

RN RNaseH

RT reverse transcriptase

SHBsAg short hepatitis B surface antigen

SIV simian immunodeficiency virus

SNP single nucleotide polymorphism

SP spacer

STHBV stork hepatitis B virus

SVM support vector machine

TDF Tenofovir

TM transmembrane region

TP terminal protein

WHV woodchuck hepatitis B virus

WMHBV woolly monkey hepatitis virus

Bibliography

Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J. J., Mayer, P., and Kawashima, E. (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res*, 28(20):E87.

Alcantara, L. C., Cassol, S., Libin, P., Deforche, K., Pybus, O. G., Van Ranst, M., GalvÃčo-Castro, B., Vandamme, A. M., and de Oliveira, T. (2009). A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. *Nucleic Acids Res*, 37(Web Server issue):W634–42.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–10.

Arauz-Ruiz, P., Norder, H., Robertson, B. H., and Magnius, L. O. (2002). Genotype H: a new Amerindian genotype of hepatitis B virus revealed in Central America. *J Gen Virol*, 83(Pt 8):2059–73.

Balabanova, Y., Gilsdorf, A., Buda, S., Burger, R., Eckmanns, T., Gärtner, B., Gross, U., Haas, W., Hamouda, O., Hübner, J., Jänisch, T., Kist, M., Kramer, M. H., Ledig, T., Mielke, M., Pulz, M., Stark, K., Suttorp, N., Ulbrich, U., Wichmann, O., and Krause, G. (2011). Communicable diseases prioritized for surveillance and epidemiological research: results of a standardized prioritization procedure in Germany, 2011. *PLoS One*, 6(10):e25691.

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

Bartenschlager, R., Junker-Niepmann, M., and Schaller, H. (1990). The P gene product of hepatitis B virus is required as a structural component for genomic RNA encapsidation. *J Virol*, 64(11):5324–32.

Bartenschlager, R. and Schaller, H. (1988). The amino-terminal domain of the hepadnaviral P-gene encodes the terminal protein (genome-linked protein) believed to prime reverse transcription. *EMBO J*, 7(13):4185–92.

Baumert, T. F., Rösler, C., Malim, M. H., and von Weizsäcker, F. (2007). Hepatitis B virus DNA is subject to extensive editing by the human deaminase APOBEC3C. *Hepatology*, 46(3):682–9.

Beale, R. C., Petersen-Mahrt, S. K., Watt, I. N., Harris, R. S., Rada, C., and Neuberger, M. S. (2004). Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: correlation with mutation spectra in vivo. *J Mol Biol*, 337(3):585–96.

Beck, J. and Nassal, M. (2007). Hepatitis B virus replication. *World J Gastroenterol*, 13(1):48–64.

Beerenwinkel, N., Günthard, H. F., Roth, V., and Metzner, K. J. (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in Microbiology*, 3(329).

Beggel, B., Münk, C., Däumer, M., Hauck, K., Häussinger, D., Lengauer, T., and Erhardt, A. (2013a). Full genome ultra-deep pyrosequencing associates G-to-A hypermutation of the hepatitis B virus genome with the natural progression of hepatitis B. *J Viral Hepat*, 20(12):882–9.

Beggel, B., Neumann-Fraune, M., Döring, M., Lawyer, G., Kaiser, R., Verheyen, J., and Lengauer, T. (2012). Genotyping hepatitis B virus dual infections using population-based sequence data. *J Gen Virol*, 93(Pt 9):1899–907.

Beggel, B., Neumann-Fraune, M., Kaiser, R., Verheyen, J., and Lengauer, T. (2013b). Inferring short-range linkage information from sequencing chromatograms. *PLoS One*, 8(12):e81687.

Bennett, R. P., Presnyak, V., Wedekind, J. E., and Smith, H. C. (2008). Nuclear Exclusion of the HIV-1 host defense factor APOBEC3G requires a novel cytoplasmic retention signal and is not dependent on RNA binding. *J Biol Chem*, 283(12):7320–7.

Benson, D. A., Karsch-Mizrachi, I., Clark, K., Lipman, D. J., Ostell, J., and Sayers, E. W. (2012). Genbank. *Nucleic Acids Res*, 40(Database issue):D48–53.

Bing, D. H., Boles, C., Rehman, F. N., Audeh, M., Belmarsh, M., Kelley, B., and Adams, C. P. (1996). Bridge amplification: a solid phase PCR system for the amplification and detection of allelic differences in single copy genes. In *Genetic Identity Conference Proceedings, Seventh International Symposium on Human Identification.*

Bishop, K. N., Holmes, R. K., and Malim, M. H. (2006). Antiviral potency of APOBEC proteins does not correlate with cytidine deamination. *J Virol*, 80(17):8450–8.

Bishop, K. N., Verma, M., Kim, E. Y., Wolinsky, S. M., and Malim, M. H. (2008). APOBEC3G inhibits elongation of HIV-1 reverse transcripts. *PLoS Pathog*, 4(12):e1000231.

Blumberg, B. S., Alter, H. J., and Visnich, S. (1965). A "new" antigen in leukemia sera. *JAMA*, 191:541–6.

Bogerd, H. P., Wiegand, H. L., Doehle, B. P., Lueders, K. K., and Cullen, B. R. (2006). APOBEC3A and APOBEC3B are potent inhibitors of LTR-retrotransposon function in human cells. *Nucleic Acids Res*, 34(1):89–95.

Bollyky, P. L., Rambaut, A., Grassly, N., and Carman, W. F. (1997). Hepatitis B virus has a recent new world evolutionary origin. *Hepatology*, 26(4):320A.

Bonvin, M., Achermann, F., Greeve, I., Stroka, D., Keogh, A., Inderbitzin, D., Candinas, D., Sommer, P., Wain-Hobson, S., Vartanian, J. P., and Greeve, J. (2006). Interferon-inducible expression of APOBEC3 editing enzymes in human hepatocytes and inhibition of hepatitis B virus replication. *Hepatology*, 43(6):1364–74.

Bonvin, M. and Greeve, J. (2007). Effects of point mutations in the cytidine deaminase domains of APOBEC3B on replication and hypermutation of hepatitis B virus in vitro. *J Gen Virol*, 88(Pt 12):3270–4.

Boser, B. E., Guyon, I., and Vapnik, V. (1992). Proceedings of the 5th Annual Workshop on Computational Learning Theory. In Haussler, D., editor, *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 144–152. ACM Press.

Böttcher, B., Vogel, M., Ploss, M., and Nassal, M. (2006). High plasticity of the hepatitis B virus capsid revealed by conformational stress. *J Mol Biol*, 356(3):812–22.

Böttcher, B., Wynne, S. A., and Crowther, R. A. (1997). Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. *Nature*, 386(6620):88–91.

Bruns, M., Miska, S., Chassot, S., and Will, H. (1998). Enhancement of hepatitis B virus infection by noninfectious subviral particles. *J Virol*, 72(2):1462–8.

Bruss, V. (1997). A short linear sequence in the pre-S domain of the large hepatitis B virus envelope protein required for virion formation. *J Virol*, 71(12):9350–7.

Buckwold, V. E., Xu, Z., Chen, M., Yen, T. S., and Ou, J. H. (1996). Effects of a naturally occurring mutation in the hepatitis B virus basal core promoter on precore gene expression and viral replication. *J Virol*, 70(9):5845–51.

Buster, E. H., Baak, B. C., Bakker, C. M., Beuers, U. H., Brouwer, J. T., Drenth, J. P., van Erpecum, K. J., van Hoek, B., Honkoop, P., Kerbert-Dreteler, M. J., Koek, G. H., van Nieuwkerk, K. M., van Soest, H., van der Spek, B. W., Tan, A. C., Vrolijk, J. M., and Janssen, H. L. (2012). The 2012 revised Dutch national guidelines for the treatment of chronic hepatitis B virus infection. *Neth J Med*, 70(8):381–5.

Buster, E. H., Flink, H. J., Cakaloglu, Y., Simon, K., Trojan, J., Tabak, F., So, T. M., Feinman, S. V., Mach, T., Akarca, U. S., Schutten, M., Tielemans, W., van Vuuren, A. J., Hansen, B. E., and Janssen, H. L. (2008). Sustained HBeAg and HBsAg loss after long-term follow-up of HBeAg-positive patients treated with peginterferon alpha-2b. *Gastroenterology*, 135(2):459–67.

Buster, E. H., Hansen, B. E., Lau, G. K., Piratvisuth, T., Zeuzem, S., Steyerberg, E. W., and Janssen, H. L. (2009). Factors that predict response of patients with hepatitis B e antigen-positive chronic hepatitis B to peginterferon-alfa. *Gastroenterology*, 137(6):2002–9.

Carman, W. F., Zanetti, A. R., Karayiannis, P., Waters, J., Manzillo, G., Tanzi, E., Zuckerman, A. J., and Thomas, H. C. (1990). Vaccine-induced escape mutant of hepatitis B virus. *Lancet*, 336(8711):325–9.

Carr, I. M., Robinson, J. I., Dimitriou, R., Markham, A. F., Morgan, A. W., and Bonthron, D. T. (2009). Inferring relative proportions of DNA variants from sequencing electropherograms. *Bioinformatics*, 25(24):3244–50.

Chain, B. M. and Myers, R. (2005). Variability and conservation in hepatitis B virus core protein. *BMC Microbiol*, 5:33.

Chaisson, M. J., Brinza, D., and Pevzner, P. A. (2009). De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res*, 19(2):336–46.

Chan, H. L., Hui, A. Y., Wong, M. L., Tse, A. M., Hung, L. C., Wong, V. W., and Sung, J. J. (2004). Genotype C hepatitis B virus infection is associated with an increased risk of hepatocellular carcinoma. *Gut*, 53(10):1494–8.

Chan, H. L., Leung, N. W., Hui, A. Y., Wong, V. W., Liew, C. T., Chim, A. M., Chan, F. K., Hung, L. C., Lee, Y. T., Tam, J. S., Lam, C. W., and Sung, J. J. (2005). A randomized, controlled trial of combination therapy for chronic hepatitis B: comparing pegylated interferon-alpha2b and lamivudine with lamivudine alone. *Ann Intern Med*, 142(4):240–50.

Chan, H. L., Thompson, A., Martinot-Peignoux, M., Piratvisuth, T., Cornberg, M., Brunetto, M. R., Tillmann, H. L., Kao, J. H., Jia, J. D., Wedemeyer, H., Locarnini, S., Janssen, H. L., and Marcellin, P. (2011). Hepatitis B surface antigen quantification: why and how to use it in 2011 - a core group report. *J Hepatol*, 55(5):1121–31.

Chang, C., Enders, G., Sprengel, R., Peters, N., Varmus, H. E., and Ganem, D. (1987). Expression of the precore region of an avian hepatitis B virus is not required for viral replication. *J Virol*, 61(10):3322–5.

Chang, L. J., Hirsch, R. C., Ganem, D., and Varmus, H. E. (1990). Effects of insertional and point mutations on the functions of the duck hepatitis B virus polymerase. *J Virol*, 64(11):5553–8.

Chauhan, R., Kazim, S. N., Kumar, M., Bhattacharjee, J., Krishnamoorthy, N., and Sarin, S. K. (2008). Identification and characterization of genotype A and D recombinant hepatitis B virus from Indian chronic HBV isolates. *World J Gastroenterol*, 14(40):6228–36.

Chelico, L., Pham, P., Calabrese, P., and Goodman, M. F. (2006). APOBEC3G DNA deaminase acts processively 3' –> 5' on single-stranded dna. *Nat Struct Mol Biol*, 13(5):392–9.

Chen, C. J., Yang, H. I., Su, J., Jen, C. L., You, S. L., Lu, S. N., Huang, G. T., Iloeje, U. H., and Group, R.-H. S. (2006). Risk of hepatocellular carcinoma across a biological gradient of serum hepatitis B virus DNA level. *JAMA*, 295(1):65–73.

Chen, D. S. (2009). Hepatitis B vaccination: The key towards elimination and eradication of hepatitis B. *J Hepatol*, 50(4):805–16.

Chen, H. S., Kaneko, S., Girones, R., Anderson, R. W., Hornbuckle, W. E., Tennant, B. C., Cote, P. J., Gerin, J. L., Purcell, R. H., and Miller, R. H. (1993). The woodchuck hepatitis virus X gene is important for establishment of virus infection in woodchucks. *J Virol*, 67(3):1218–26.

Chen, M. T., Billaud, J. N., Sällberg, M., Guidotti, L. G., Chisari, F. V., Jones, J., Hughes, J., and Milich, D. R. (2004). A function of the hepatitis B virus precore protein is to regulate the immune response to the core antigen. *Proc Natl Acad Sci USA*, 101(41):14913–8.

Chiou, H. L., Lee, T. S., Kuo, J., Mau, Y. C., and Ho, M. S. (1997). Altered antigenicity of 'a' determinant variants of hepatitis B virus. *J Gen Virol*, 78 (Pt 10):2639–45.

Chiu, Y. L. and Greene, W. C. (2008). The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. *Annu Rev Immunol*, 26:317–53.

Chu, C. J., Keeffe, E. B., Han, S. H., Perrillo, R. P., Min, A. D., Soldevila-Pico, C., Carey, W., Brown, R. S., Luketic, V. A., Terrault, N., Lok, A. S., and Group, U. H. E. S. (2003). Prevalence of hbv precore/core promoter variants in the united states. *Hepatology*, 38(3):619–28.

Chu, C. M., Hung, S. J., Lin, J., Tai, D. I., and Liaw, Y. F. (2004). Natural history of hepatitis B e antigen to antibody seroconversion in patients with normal serum aminotransferase levels. *Am J Med*, 116(12):829–34.

Conway, J. F., Cheng, N., Zlotnick, A., Wingfield, P. T., Stahl, S. J., and Steven, A. C. (1997). Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy. *Nature*, 386(6620):91–4.

Conway, J. F., Watts, N. R., Belnap, D. M., Cheng, N., Stahl, S. J., Wingfield, P. T., and Steven, A. C. (2003). Characterization of a conformational epitope on hepatitis B virus core antigen and quasiequivalent variations in antibody binding. *J Virol*, 77(11):6466–73.

Cooksley, W. G. (2010). Do we need to determine viral genotype in treating chronic hepatitis B? *J Viral Hepat*, 17(9):601–10.

Cornberg, M., Protzer, U., Petersen, J., Wedemeyer, H., Berg, T., Jilg, W., Erhardt, A., Wirth, S., Sarrazin, C., Dollinger, M. M., Schirmacher, P., Dathe, K., Kopp, I. B., Zeuzem, S., Gerlich, W. H., and Manns, M. P. (2011). Prophylaxis, diagnosis and therapy of hepatitis B virus infection - the German guideline. *Z Gastroenterol*, 49(7):871–930.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Mach. Learn.*, 20(3):273–297.

Cox, R. T. (1946). Probability, Frequency and Reasonable Expectation. *American Journal of Physics*, 14(1):1–13.

Crowther, R. A., Kiselev, N. A., Böttcher, B., Berriman, J. A., Borisova, G. P., Ose, V., and Pumpens, P. (1994). Three-dimensional structure of hepatitis B virus core particles determined by electron cryomicroscopy. *Cell*, 77(6):943–50.

Cui, C., Shi, J., Hui, L., Xi, H., Zhuoma, Quni, Tsedan, and Hu, G. (2002). The dominant hepatitis B virus genotype identified in Tibet is a C/D hybrid. *J Gen Virol*, 83(Pt 11):2773–7.

Datta, S., Panigrahi, R., Biswas, A., Chandra, P. K., Banerjee, A., Mahapatra, P. K., Panda, C. K., Chakrabarti, S., Bhattacharya, S. K., Biswas, K., and Chakravarty, R. (2009). Genetic characterization of hepatitis B virus in peripheral blood leukocytes: evidence for selection and compartmentalization of viral variants with the immune escape G145R mutation. *J Virol*, 83(19):9983–92.

de Finetti, B. (1974). *Theory of Probability: A Critical Introductory Treatment*. Wiley.

Delebecque, F., Suspène, R., Calattini, S., Casartelli, N., Saïb, A., Froment, A., Wain-Hobson, S., Gessain, A., Vartanian, J. P., and Schwartz, O. (2006). Restriction of foamy viruses by APOBEC cytidine deaminases. *J Virol*, 80(2):605–14.

Devesa, M. and Pujol, F. H. (2007). Hepatitis B virus genetic diversity in Latin America. *Virus Res*, 127(2):177–84.

Dmitriev, D. A. and Rakitov, R. A. (2008). Decoding of superimposed traces produced by direct sequencing of heterozygous indels. *PLoS Comput Biol*, 4(7):e1000113.

Döring, M. (2011). Web Implementation of the HBV Dual Infection Model. Bachelor's thesis, Saarland University.

Durbin, R., Eddy, S., A., K., and G., M. (1998). *Biological sequence analysis*. Cambridge University Press.

Eble, B. E., Lingappa, V. R., and Ganem, D. (1990). The N-terminal (pre-S2) domain of a hepatitis B virus surface glycoprotein is translocated across membranes by downstream signal sequences. *J Virol*, 64(3):1414–9.

Eble, B. E., MacRae, D. R., Lingappa, V. R., and Ganem, D. (1987). Multiple topogenic sequences determine the transmembrane orientation of the hepatitis B surface antigen. *Mol Cell Biol*, 7(10):3591–601.

European Association for the Study of the Liver (2012). EASL clinical practice guidelines: Management of chronic hepatitis B virus infection. *J Hepatol*, 57(1):167–85.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*, 8(3):175–85.

Fensterl, V. and Sen, G. C. (2009). Interferons and viral infections. *Biofactors*, 35(1):14–20.

Fields, B. N., Knipe, D. M., and Howley, P. M. (2007). *Fields' virology*. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, 5th edition.

Flood, E. M., Tang, F., Horvath, M. M., Pertsemlidis, A., and Garner, H. R. (2002). SNPCEQer: detecting SNPs in sequences generated by the Beckman CEQ2000 DNA Analysis System. *Biotechniques*, 33(4):814, 816, 818–20 passim.

Flot, J.-F. (2007). CHAMPURU 1.0: a computer software for unraveling mixtures of two DNA sequences of unequal lengths. *Molecular Ecology Notes*, 7:974–977.

Flot, J.-F., Tillier, A., Samadi, S., and Tillier, S. (2006). Phase determination from direct sequencing of length-variable DNA regions. *Molecular Ecology Notes*, 6:627–630.

Gabuzda, D. H., Lawrence, K., Langhoff, E., Terwilliger, E., Dorfman, T., Haseltine, W. A., and Sodroski, J. (1992). Role of vif in replication of human immunodeficiency virus type 1 in CD4+ T lymphocytes. *J Virol*, 66(11):6489–95.

Gale, C. V., Myers, R., Tedder, R. S., Williams, I. G., and Kellam, P. (2004). Development of a novel human immunodeficiency virus type 1 subtyping tool, Subtype Analyzer (STAR): analysis of subtype distribution in London. *AIDS Res Hum Retroviruses*, 20(5):457–64.

Gallina, A., Bonelli, F., Zentilin, L., Rindi, G., Muttini, M., and Milanesi, G. (1989). A recombinant hepatitis B core antigen polypeptide with the protamine-like domain deleted self-assembles into capsid particles but fails to bind nucleic acids. *J Virol*, 63(11):4645–52.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M., and Hahn, B. H. (1999). Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature*, 397(6718):436–41.

Gao, F., Yue, L., White, A. T., Pappas, P. G., Barchue, J., Hanson, A. P., Greene, B. M., Sharp, P. M., Shaw, G. M., and Hahn, B. H. (1992). Human infection by genetically diverse SIVSM-related HIV-2 in west Africa. *Nature*, 358(6386):495–9.

Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., and Beerenwinkel, N. (2012). Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun*, 3:811.

Gilbert, R. J., Beales, L., Blond, D., Simon, M. N., Lin, B. Y., Chisari, F. V., Stuart, D. I., and Rowlands, D. J. (2005). Hepatitis B small surface antigen particles are octahedral. *Proc Natl Acad Sci USA*, 102(41):14783–8.

Gonzalez, M. C., Suspène, R., Henry, M., Guétard, D., Wain-Hobson, S., and Vartanian, J. P. (2009). Human APOBEC1 cytidine deaminase edits HBV DNA. *Retrovirology*, 6:96.

Good, I. (1950). *Probability and the Weighing of Evidence*. Charles Griffin.

Günther, S., Sommer, G., Plikat, U., Iwanska, A., Wain-Hobson, S., Will, H., and Meyerhans, A. (1997). Naturally occurring hepatitis B virus genomes bearing the hallmarks of retroviral G-->A hypermutation. *Virology*, 235(1):104–8.

Guttman, A., Cohen, A. S., Heiger, D. N., and Karger, B. L. (1990). Analytical and Micropreparative Ultrahigh Resolution of Oligonucleotides by Polyacrylamide Gel High-Performance Capillary Electrophoresis. *Anal Chem*, 62(2):137–141.

Guyon, I. and Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.*, 3:1157–1182.

Hackshaw, A. (2008). Small studies: strengths and limitations. *Eur Respir J*, 32(5):1141–3.

Hadziyannis, S. J., Tassopoulos, N. C., Heathcote, E. J., Chang, T. T., Kitis, G., Rizzetto, M., Marcellin, P., Lim, S. G., Goodman, Z., Ma, J., Brosgart, C. L., Borroto-Esoda, K., Arterburn, S., Chuck, S. L., and Group, A. D. . S. (2006). Long-term therapy with adefovir dipivoxil for HBeAg-negative chronic hepatitis B for up to 5 years. *Gastroenterology*, 131(6):1743–51.

Haines, K. M. and Loeb, D. D. (2007). The sequence of the RNA primer and the DNA template influence the initiation of plus-strand DNA synthesis in hepatitis B virus. *J Mol Biol*, 370(3):471–80.

Hannoun, C., Krogsgaard, K., Horal, P., Lindh, M., and group, I. t. (2002). Genotype mixtures of hepatitis B virus in patients treated with interferon. *J Infect Dis*, 186(6):752–9.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer.

Heathcote, E. J., Marcellin, P., Buti, M., Gane, E., De Man, R. A., Krastev, Z., Germanidis, G., Lee, S. S., Flisiak, R., Kaita, K., Manns, M., Kotzev, I., Tchernev, K., Buggisch, P., Weilert, F., Kurdas, O. O., Shiffman, M. L., Trinh, H., Gurel, S., Snow-Lampart, A., Borroto-Esoda, K., Mondou, E., Anderson, J., Sorbel, J., and Rousseau, F. (2011). Three-year efficacy and safety of tenofovir disoproxil fumarate treatment for chronic hepatitis B. *Gastroenterology*, 140(1):132–43.

Henry, M., Guétard, D., Suspène, R., Rusniok, C., Wain-Hobson, S., and Vartanian, J. P. (2009). Genetic editing of HBV DNA by monodomain human APOBEC3 cytidine deaminases and the recombinant nature of APOBEC3G. *PLoS One*, 4(1):e4277.

Hirsch, V. M., Olmsted, R. A., Murphey-Corb, M., Purcell, R. H., and Johnson, P. R. (1989). An african primate lentivirus (SIVsm) closely related to HIV-2. *Nature*, 339(6223):389–92.

Hoofnagle, J. H., Di Bisceglie, A. M., Waggoner, J. G., and Park, Y. (1993). Interferon alfa for patients with clinically apparent cirrhosis due to chronic hepatitis B. *Gastroenterology*, 104(4):1116–21.

Hoofnagle, J. H., Doo, E., Liang, T. J., Fleischer, R., and Lok, A. S. (2007). Management of hepatitis B: summary of a clinical research workshop. *Hepatology*, 45(4):1056–75.

Hsieh, C. C., Tzonou, A., Zavitsanos, X., Kaklamani, E., Lan, S. J., and Trichopoulos, D. (1992). Age at first establishment of chronic hepatitis B virus infection and hepatocellular carcinoma risk. a birth order study. *Am J Epidemiol*, 136(9):1115–21.

Hui, C. K., Leung, N., Yuen, S. T., Zhang, H. Y., Leung, K. W., Lu, L., Cheung, S. K., Wong, W. M., Lau, G. K., and Group, H. K. L. F. S. (2007). Natural history and

disease progression in Chinese chronic hepatitis B patients in immune-tolerant phase. *Hepatology*, 46(2):395–401.

Huovila, A. P., Eder, A. M., and Fuller, S. D. (1992). Hepatitis b surface antigen assembles in a post-er, pre-golgi compartment. *J Cell Biol*, 118(6):1305–20.

Hussain, M., Chu, C. J., Sablon, E., and Lok, A. S. (2003). Rapid and sensitive assays for determination of hepatitis B virus (HBV) genotypes and detection of HBV precore and core promoter variants. *J Clin Microbiol*, 41(8):3699–705.

Iannacone, M., Sitia, G., Ruggeri, Z. M., and Guidotti, L. G. (2007). HBV pathogenesis in animal models: recent advances on the role of platelets. *J Hepatol*, 46(4):719–26.

Iloeje, U. H., Yang, H. I., Su, J., Jen, C. L., You, S. L., Chen, C. J., and The Risk Evaluation of Viral Load Elevation and Associated Liver Disease/Cancer-In HBV (the REVEAL-HBV) Study Group (2006). Predicting cirrhosis risk based on the level of circulating hepatitis B viral load. *Gastroenterology*, 130(3):678–86.

Isaacs, A. and Lindenmann, J. (1957). Virus interference: I. The interferon. *CA Cancer J Clin*, 38(5):280–90.

Janahi, E. M. and McGarvey, M. J. (2013). The inhibition of hepatitis B virus by APOBEC cytidine deaminases. *J Viral Hepat*, 20(12):821–896.

Janitz, M. (2011). *Next-Generation Genome Sequencing: Towards Personalized Medicine*. John Wiley & Sons.

Janssen, H. L., van Zonneveld, M., Senturk, H., Zeuzem, S., Akarca, U. S., Cakaloglu, Y., Simon, C., So, T. M., Gerken, G., de Man, R. A., Niesters, H. G., Zondervan, P., Hansen, B., and Schalm, S. W. (2005). Pegylated interferon alfa-2b alone or in combination with lamivudine for HBeAg-positive chronic hepatitis B: a randomised trial. *Lancet*, 365(9454):123–9.

Japkowicz, N. and Stephen, S. (2002). The Class Imbalance Problem: A Systematic Study. *Intell. Data Anal.*, 6(5):429–449.

Jarmuz, A., Chester, A., Bayliss, J., Gisbourne, J., Dunham, I., Scott, J., and Navaratnam, N. (2002). An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. *Genomics*, 79(3):285–96.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK, New York.

Jost, S., Turelli, P., Mangeat, B., Protzer, U., and Trono, D. (2007). Induction of antiviral cytidine deaminases does not explain the inhibition of hepatitis B virus replication by interferons. *J Virol*, 81(19):10588–96.

Ju, J., Ruan, C., Fuller, C. W., Glazer, A. N., and Mathies, R. A. (1995). Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis. *Proc Natl Acad Sci USA*, 92(10):4347–51.

Kamatani, Y., Wattanapokayakit, S., Ochi, H., Kawaguchi, T., Takahashi, A., Hosono, N., Kubo, M., Tsunoda, T., Kamatani, N., Kumada, H., Puseenam, A., Sura, T., Daigo, Y., Chayama, K., Chantratita, W., Nakamura, Y., and Matsuda, K. (2009). A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. *Nat Genet*, 41(5):591–5.

Kao, J. H., Chen, P. J., Lai, M. Y., and Chen, D. S. (2000). Hepatitis B genotypes correlate with clinical outcomes in patients with chronic hepatitis B. *Gastroenterology*, 118(3):554–9.

Kao, J. H., Chen, P. J., Lai, M. Y., and Chen, D. S. (2001). Acute exacerbations of chronic hepatitis B are rarely associated with superinfection of hepatitis B virus. *Hepatology*, 34(4 Pt 1):817–23.

Karush, W. (1939). Minima of Functions of Several Variables with Inequalities as Side Constraints. Master's thesis, Deptartment of Mathematics, University of Chicago.

Kato, H., Orito, E., Gish, R. G., Sugauchi, F., Suzuki, S., Ueda, R., Miyakawa, Y., and Mizokami, M. (2002). Characteristics of hepatitis B virus isolates of genotype G and their phylogenetic differences from the other six genotypes (A through F). *J Virol*, 76(12):6131–7.

Kato, H., Orito, E., Sugauchi, F., Ueda, R., Koshizaka, T., Yanaka, S., Gish, R. G., Kurbanov, F., Ruzibakiev, R., Kramvis, A., Kew, M. C., Ahmad, N., Khan, M., Usuda, S., Miyakawa, Y., and Mizokami, M. (2003). Frequent coinfection with hepatitis B virus strains of distinct genotypes detected by hybridization with type-specific probes immobilized on a solid-phase support. *J Virol Methods*, 110(1):29–35.

Keith, C. S., Hoang, D. O., Barrett, B. M., Feigelman, B., Nelson, M. C., Thai, H., and Baysdorfer, C. (1993). Partial sequence analysis of 130 randomly selected maize cDNA clones. *Plant Physiol*, 101(1):329–32.

Kinomoto, M., Kanno, T., Shimura, M., Ishizaka, Y., Kojima, A., Kurata, T., Sata, T., and Tokunaga, K. (2007). All APOBEC3 family proteins differentially inhibit LINE-1 retrotransposition. *Nucleic Acids Res*, 35(9):2955–64.

Köck, J. and Blum, H. E. (2008). Hypermutation of hepatitis B virus genomes by APOBEC3G, APOBEC3C and APOBEC3H. *J Gen Virol*, 89(Pt 5):1184–91.

Kuhn, H. and Tucker, A. (1951). Nonlinear programming. In Neyman, J., editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492. University of California Press, Berkeley, California.

Kumar, M., Jung, S. Y., Hodgson, A. J., Madden, C. R., Qin, J., and Slagle, B. L. (2011). Hepatitis B virus regulatory HBx protein binds to adaptor protein IPS-1 and inhibits the activation of beta interferon. *J Virol*, 85(2):987–95.

Kumar, V., Jayasuryan, N., and Kumar, R. (1996). A truncated mutant (residues 58-140) of the hepatitis B virus X protein retains transactivation function. *Proc Natl Acad Sci USA*, 93(11):5647–52.

Kwok, P. Y., Carlson, C., Yager, T. D., Ankener, W., and Nickerson, D. A. (1994). Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics*, 23(1):138–44.

Kwon, H. and Lok, A. S. (2011). Hepatitis B therapy. *Nat Rev Gastroenterol Hepatol*, 8(5):275–84.

Lackey, L., Demorest, Z. L., Land, A. M., Hultquist, J. F., Brown, W. L., and Harris, R. S. (2012). APOBEC3B and AID have similar nuclear import mechanisms. *J Mol Biol*, 419(5):301–14.

Lai, C. L., Gane, E., Liaw, Y. F., Hsu, C. W., Thongsawat, S., Wang, Y., Chen, Y., Heathcote, E. J., Rasenack, J., Bzowej, N., Naoumov, N. V., Di Bisceglie, A. M., Zeuzem, S., Moon, Y. M., Goodman, Z., Chao, G., Constance, B. F., Brown, N. A., and Group, G. S. (2007). Telbivudine versus lamivudine in patients with chronic hepatitis B. *N Engl J Med*, 357(25):2576–88.

Lai, C. L., Shouval, D., Lok, A. S., Chang, T. T., Cheinquer, H., Goodman, Z., DeHertogh, D., Wilber, R., Zink, R. C., Cross, A., Colonno, R., Fernandes, L., and Group, B. A. S. (2006). Entecavir versus lamivudine for patients with HBeAg-negative chronic hepatitis B. *N Engl J Med*, 354(10):1011–20.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lanford, R. E., Notvall, L., Lee, H., and Beames, B. (1997). Transcomplementation of nucleotide priming and reverse transcription between independently expressed TP and RT domains of the hepatitis B virus reverse transcriptase. *J Virol*, 71(4):2996–3004.

Lange, C. M., Bojunga, J., Hofmann, W. P., Wunder, K., Mihm, U., Zeuzem, S., and Sarrazin, C. (2009). Severe lactic acidosis during treatment of chronic hepatitis B with entecavir in patients with impaired liver function. *Hepatology*, 50(6):2001–6.

Lau, G. K., Piratvisuth, T., Luo, K. X., Marcellin, P., Thongsawat, S., Cooksley, G., Gane, E., Fried, M. W., Chow, W. C., Paik, S. W., Chang, W. Y., Berg, T., Flisiak, R., McCloud, P., and Pluck, N. (2005). Peginterferon Alfa-2a, lamivudine, and the combination for HBeAg-positive chronic hepatitis B. *N Engl J Med*, 352(26):2682–95.

Lavanchy, D. (2004). Hepatitis B virus epidemiology, disease burden, treatment, and current and emerging prevention and control measures. *J Viral Hepat*, 11(2):97–107.

Lee, H. W., Lee, H. J., Hwang, J. S., Sohn, J. H., Jang, J. Y., Han, K. J., Park, J. Y., Kim, d. Y., Ahn, S. H., Paik, Y. H., Lee, C. K., Lee, K. S., Chon, C. Y., and Han, K. H. (2010). Lamivudine maintenance beyond one year after HBeAg seroconversion is a major factor for sustained virologic response in HBeAg-positive chronic hepatitis B. *Hepatology*, 51(2):415–21.

Lee, K. M., Kim, Y. S., Ko, Y. Y., Yoo, B. M., Lee, K. J., Kim, J. H., Hahm, K. B., and Cho, S. W. (2001). Emergence of vaccine-induced escape mutant of hepatitis B virus with multiple surface gene mutations in a Korean child. *J Korean Med Sci*, 16(3):359–62.

Lee, L. G., Connell, C. R., Woo, S. L., Cheng, R. D., McArdle, B. F., Fuller, C. W., Halloran, N. D., and Wilson, R. K. (1992). DNA sequencing with dye-labeled terminators and T7 DNA polymerase: effect of dyes and dNTPs on incorporation of dye-terminators and probability analysis of termination fragments. *Nucleic Acids Res*, 20(10):2471–83.

Lee, W. M. (1993). Acute liver failure. *N Engl J Med*, 329(25):1862–72.

Leistner, C. M., Gruen-Bernhard, S., and Glebe, D. (2008). Role of glycosaminoglycans for binding and infection of hepatitis B virus. *Cell Microbiol*, 10(1):122–33.

Lemon, S. M. and Thomas, D. L. (1997). Vaccines to prevent viral hepatitis. *N Engl J Med*, 336(3):196–204.

Lengauer, T., Sander, O., Sierra, S., Thielen, A., and Kaiser, R. (2007). Bioinformatics prediction of HIV coreceptor usage. *Nat Biotechnol*, 25(12):1407–10.

Lengauer, T. and Sing, T. (2006). Bioinformatics-assisted anti-HIV therapy. *Nat Rev Microbiol*, 4(10):790–7.

Lewellyn, E. B. and Loeb, D. D. (2007). Base pairing between cis-acting sequences contributes to template switching during plus-strand DNA synthesis in human hepatitis B virus. *J Virol*, 81(12):6207–15.

Li, M. M. and Emerman, M. (2011). Polymorphism in human APOBEC3H affects a phenotype dominant for subcellular localization and antiviral activity. *J Virol*, 85(16):8197–207.

Li, Y., Mitaxov, V., and Waksman, G. (1999). Structure-based design of Taq DNA polymerases with improved properties of dideoxynucleotide incorporation. *Proc Natl Acad Sci USA*, 96(17):9491–6.

Liao, W., Hong, S. H., Chan, B. H., Rudolph, F. B., Clark, S. C., and Chan, L. (1999). APOBEC-2, a cardiac- and skeletal muscle-specific member of the cytidine deaminase supergene family. *Biochem Biophys Res Commun*, 260(2):398–404.

Liao, W. and Ou, J. H. (1995). Phosphorylation and nuclear localization of the hepatitis B virus core protein: significance of serine in the three repeated SPRRR motifs. *J Virol*, 69(2):1025–9.

Liaw, Y. F., Leung, N., Kao, J. H., Piratvisuth, T., Gane, E., Han, K. H., Guan, R., Lau, G. K., and Locarnini, S. (2008). Asian-pacific consensus statement on the management of chronic hepatitis B: a 2008 update. *Hepatol Int*, 2(3):263–83.

Lin, C. L. and Kao, J. H. (2011). The clinical implications of hepatitis B virus genotype: Recent advances. *J Gastroenterol Hepatol*, 26 Suppl 1:123–30.

Lindh, M., Uhnoo, I., Bläckberg, J., Duberg, A. S., Friman, S., Fischler, B., Karlström, O., Norkrans, G., Reichard, O., Sangfeldt, P., Söderström, A., Sönnerborg, A., Weiland, O., Wejstål, R., and Wiström, J. (2008). Treatment of chronic hepatitis B infection: an update of Swedish recommendations. *Scand J Infect Dis*, 40(6-7):436–50.

Lindley, D. V. (1982). Scoring Rules and the Inevitability of Probability. *International Statistical Review / Revue Internationale de Statistique*, 50(1):pp. 1–11.

Livingston, S. E., Simonetti, J. P., McMahon, B. J., Bulkow, L. R., Hurlburt, K. J., Homan, C. E., Snowball, M. M., Cagle, H. H., Williams, J. L., and Chulanov, V. P. (2007). Hepatitis B virus genotypes in Alaska Native people with hepatocellular carcinoma: preponderance of genotype F. *J Infect Dis*, 195(1):5–11.

Locarnini, S., Littlejohn, M., Aziz, M. N., and Yuen, L. (2013). Possible origins and evolution of the hepatitis B virus (HBV). *Semin Cancer Biol*.

Lok, A. S. and McMahon, B. J. (2001). Chronic hepatitis B. *Hepatology*, 34(6):1225–41.

Lok, A. S. and McMahon, B. J. (2009). Chronic hepatitis B: update 2009. *Hepatology*, 50(3):661–2.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., and Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol*, 30(5):434–9.

Luckey, J. A., Drossman, H., Kostichka, A. J., Mead, D. A., D'Cunha, J., Norris, T. B., and Smith, L. M. (1990). High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Res*, 18(15):4417–21.

Lycke, E. (1976). 1976 Nobel prize winner in physiology or medicine: Discoveries of new factors for etiology and distribution of infectious diseases. *Lakartidningen*, 73(44):3743–6.

Mahoney, F. J. (1999). Update on diagnosis, management, and prevention of hepatitis B virus infection. *Clin Microbiol Rev*, 12(2):351–66.

Mallory, M. A., Page, S. R., and Hillyard, D. R. (2011). Development and validation of a hepatitis B virus DNA sequencing assay for assessment of antiviral resistance, viral genotype and surface antigen mutation status. *J Virol Methods*, 177(1):31–7.

Manion, M., Ni, S., Hulce, D., and Liu, J. (2009). DNA Mutation and Methylation Quantification from Sanger Sequencing Traces with Mutation Surveyor Software. http://www.softgenetics.com/.

Marcellin, P., Buti, M., Krastev, Z., Gurel, S., Di Bisceglie, A. M., Odin, J. A., Dusheiko, G. M., Heathcote, E. J., Borroto-Esoda, K., Coombs, D. H., Mondou, E., and Anderson, J. (2010). Continued efficacy and safety through 4 years of tenofovir disoproxil fumarate (TDF) treatment in HBeAg-negative patients with chronic hepatitis B (study 102). In *Hepatology*, volume 52, pages 555A–556A.

Marcellin, P., Chang, T. T., Lim, S. G., Sievert, W., Tong, M., Arterburn, S., Borroto-Esoda, K., Frederick, D., and Rousseau, F. (2008a). Long-term efficacy and safety of adefovir dipivoxil for the treatment of hepatitis B e antigen-positive chronic hepatitis B. *Hepatology*, 48(3):750–8.

Marcellin, P., Gane, E., Buti, M., Afdhal, N., Sievert, W., Jacobson, I. M., Washington, M. K., Germanidis, G., Flaherty, J. F., Schall, R. A., Bornstein, J. D., Kitrinos, K. M., Subramanian, G. M., McHutchison, J. G., and Heathcote, E. J. (2013). Regression of cirrhosis during treatment with tenofovir disoproxil fumarate for chronic hepatitis B: a 5-year open-label follow-up study. *Lancet*, 381(9865):468–75.

Marcellin, P., Heathcote, E. J., Buti, M., Gane, E., de Man, R. A., Krastev, Z., Germanidis, G., Lee, S. S., Flisiak, R., Kaita, K., Manns, M., Kotzev, I., Tchernev, K., Buggisch, P., Weilert, F., Kurdas, O. O., Shiffman, M. L., Trinh, H., Washington, M. K., Sorbel, J., Anderson, J., Snow-Lampart, A., Mondou, E., Quinn, J., and Rousseau, F. (2008b). Tenofovir disoproxil fumarate versus adefovir dipivoxil for chronic hepatitis B. *N Engl J Med*, 359(23):2442–55.

Marcellin, P., Lau, G. K., Bonino, F., Farci, P., Hadziyannis, S., Jin, R., Lu, Z. M., Piratvisuth, T., Germanidis, G., Yurdaydin, C., Diago, M., Gurel, S., Lai, M. Y., Button, P., Pluck, N., and Peginterferon Alfa-2a HBeAg-Negative Chronic Hepatitis B Study Group (2004). Peginterferon alfa-2a alone, lamivudine alone, and the two in combination in patients with HBeAg-negative chronic hepatitis B. *N Engl J Med*, 351(12):1206–17.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, 9:387–402.

Margolis, H. S. (1993). Prevention of acute and chronic liver disease through immunization: hepatitis B and beyond. *J Infect Dis*, 168(1):9–14.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B.,

McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–80.

Marsh, M. and Helenius, A. (2006). Virus entry: open sesame. *Cell*, 124(4):729–40.

Martin-Vilchez, S., Lara-Pezzi, E., Trapero-Marugán, M., Moreno-Otero, R., and Sanz-Cameno, P. (2011). The molecular and pathophysiological implications of hepatitis B X antigen in chronic hepatitis B virus infection. *Rev Med Virol*.

McAleer, W. J., Buynak, E. B., Maigetter, R. Z., Wampler, D. E., Miller, W. J., and Hilleman, M. R. (1984). Human hepatitis B vaccine from recombinant yeast. *Nature*, 307(5947):178–80.

McMahon, B. J. (2009). The natural history of chronic hepatitis B virus infection. *Hepatology*, 49(5 Suppl):S45–55.

McMahon, B. J., Holck, P., Bulkow, L., and Snowball, M. (2001). Serologic and clinical outcomes of 1536 Alaska Natives chronically infected with hepatitis B virus. *Ann Intern Med*, 135(9):759–68.

Mehta, A., Kinter, M. T., Sherman, N. E., and Driscoll, D. M. (2000). Molecular cloning of apobec-1 complementation factor, a novel RNA-binding protein involved in the editing of apolipoprotein B mRNA. *Mol Cell Biol*, 20(5):1846–54.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46.

Metzker, M. L., Lu, J., and Gibbs, R. A. (1996). Electrophoretically uniform fluorescent dyes for automated DNA sequencing. *Science*, 271(5254):1420–2.

Michitaka, K., Horiike, N., Chen, Y., Duong, T. N., Matsuura, K., Tokumoto, Y., Hiasa, Y., Akbar, F. S., and Onji, M. (2005). Co-infection with hepatitis B virus genotype D and other genotypes in Western Japan. *Intervirology*, 48(4):262–7.

Milich, D. R., Jones, J. E., Hughes, J. L., Price, J., Raney, A. K., and McLachlan, A. (1990). Is a function of the secreted hepatitis B e antigen to induce immunologic tolerance *in utero*? *Proc Natl Acad Sci USA*, 87(17):6599–603.

Morozov, V., Pisareva, M., and Groudinin, M. (2000). Homologous recombination between different genotypes of hepatitis B virus. *Gene*, 260(1-2):55–65.

Munch, K. and Krogh, A. (2006). Automatic generation of gene finders for eukaryotic species. *BMC Bioinformatics*, 7:263.

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell*, 102(5):553–63.

Muramatsu, M., Sankaranand, V. S., Anant, S., Sugai, M., Kinoshita, K., Davidson, N. O., and Honjo, T. (1999). Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. *J Biol Chem*, 274(26):18470–6.

Murray, V. (1989). Improved double-stranded DNA sequencing using the linear polymerase chain reaction. *Nucleic Acids Res*, 17(21):8889.

Myers, R., Clark, C., Khan, A., Kellam, P., and Tedder, R. (2006). Genotyping hepatitis B virus from whole- and sub-genomic fragments using position-specific scoring matrices in HBV STAR. *J Gen Virol*, 87(Pt 6):1459–64.

Myers, R. E., Gale, C. V., Harrison, A., Takeuchi, Y., and Kellam, P. (2005). A statistical model for HIV-1 sequence classification using the subtype analyser (STAR). *Bioinformatics*, 21(17):3535–40.

Naito, H., Hayashi, S., and Abe, K. (2001). Rapid and specific genotyping system for hepatitis B virus corresponding to six major genotypes by pcr using type-specific primers. *J Clin Microbiol*, 39(1):362–4.

Nassal, M. (2008). Hepatitis B viruses: reverse transcription a different way. *Virus Res*, 134(1-2):235–49.

Naumann, H., Schaefer, S., Yoshida, C. F., Gaspar, A. M., Repp, R., and Gerlich, W. H. (1993). Identification of a new hepatitis B virus (HBV) genotype from Brazil that expresses HBV surface antigen subtype adw4. *J Gen Virol*, 74 (Pt 8):1627–32.

Newbold, J. E., Xin, H., Tencza, M., Sherman, G., Dean, J., Bowden, S., and Locarnini, S. (1995). The covalently closed duplex form of the hepadnavirus genome exists in situ as a heterogeneous population of viral minichromosomes. *J Virol*, 69(6):3350–7.

Nguyen, D. H., Gummuluru, S., and Hu, J. (2007). Deamination-independent inhibition of hepatitis B virus reverse transcription by APOBEC3G. *J Virol*, 81(9):4465–72.

Ni, Y., Lempp, F. A., Mehrle, S., Nkongolo, S., Kaufman, C., FÃď lth, M., Stindt, J., Königer, C., Nassal, M., Kubitz, R., Sültmann, H., and Urban, S. (2014). Hepatitis B and D viruses exploit sodium taurocholate co-transporting polypeptide for species-specific entry into hepatocytes. *Gastroenterology*, 146(4):1070–83.

Nickerson, D. A., Tobe, V. O., and Taylor, S. L. (1997). PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res*, 25(14):2745–51.

Niederau, C., Heintges, T., Lange, S., Goldmann, G., Niederau, C. M., Mohr, L., and Häussinger, D. (1996). Long-term follow-up of HBeAg-positive patients treated with interferon alfa for chronic hepatitis B. *N Engl J Med*, 334(22):1422–7.

Noguchi, C., Hiraga, N., Mori, N., Tsuge, M., Imamura, M., Takahashi, S., Fujimoto, Y., Ochi, H., Abe, H., Maekawa, T., Yatsuji, H., Shirakawa, K., Takaori-Kondo, A., and Chayama, K. (2007). Dual effect of APOBEC3G on hepatitis B virus. *J Gen Virol*, 88(Pt 2):432–40.

Noguchi, C., Imamura, M., Tsuge, M., Hiraga, N., Mori, N., Miki, D., Kimura, T., Takahashi, S., Fujimoto, Y., Ochi, H., Abe, H., Maekawa, T., Tateno, C., Yoshizato, K., and Chayama, K. (2009). G-to-A hypermutation in hepatitis B virus (HBV) and clinical course of patients with chronic HBV infection. *J Infect Dis*, 199(11):1599–607.

Noguchi, C., Ishino, H., Tsuge, M., Fujimoto, Y., Imamura, M., Takahashi, S., and Chayama, K. (2005). G to A hypermutation of hepatitis B virus. *Hepatology*, 41(3):626–33.

Norder, H., Couroucé, A. M., Coursaget, P., Echevarria, J. M., Lee, S. D., Mushahwar, I. K., Robertson, B. H., Locarnini, S., and Magnius, L. O. (2004). Genetic diversity of hepatitis B virus strains derived worldwide: genotypes, subgenotypes, and HBsAg subtypes. *Intervirology*, 47(6):289–309.

Norder, H., Couroucé, A. M., and Magnius, L. O. (1994). Complete genomes, phylogenetic relatedness, and structural proteins of six strains of the hepatitis B virus, four of which represent two new genotypes. *Virology*, 198(2):489–503.

Ogawa, M., Hasegawa, K., Naritomi, T., Torii, N., and Hayashi, N. (2002). Clinical features and viral sequences of various genotypes of hepatitis B virus compared among patients with acute hepatitis B. *Hepatol Res*, 23(3):167–177.

Okamoto, H., Tsuda, F., Akahane, Y., Sugai, Y., Yoshiba, M., Moriyama, K., Tanaka, T., Miyakawa, Y., and Mayumi, M. (1994). Hepatitis B virus with mutations in the core promoter for an e antigen-negative phenotype in carriers with antibody to e antigen. *J Virol*, 68(12):8102–10.

Okamoto, H., Tsuda, F., Sakugawa, H., Sastrosoewignjo, R. I., Imai, M., Miyakawa, Y., and Mayumi, M. (1988). Typing hepatitis B virus by homology in nucleotide sequence: comparison of surface antigen subtypes. *J Gen Virol*, 69 (Pt 10):2575–83.

Olinger, C. M., Jutavijittum, P., Hübschen, J. M., Yousukh, A., Samountry, B., Thammavong, T., Toriyama, K., and Muller, C. P. (2008). Possible new hepatitis B virus genotype, southeast Asia. *Emerg Infect Dis*, 14(11):1777–80.

Osiowy, C. and Giles, E. (2003). Evaluation of the INNO-LiPA HBV genotyping assay for determination of hepatitis B virus genotype. *J Clin Microbiol*, 41(12):5473–7.

Osiowy, C., Gordon, D., Borlang, J., Giles, E., and Villeneuve, J. P. (2008). Hepatitis B virus genotype G epidemiology and co-infection with genotype A in Canada. *J Gen Virol*, 89(Pt 12):3009–15.

Owiredu, W. K., Kramvis, A., and Kew, M. C. (2001). Molecular analysis of hepatitis B virus genomes isolated from black African patients with fulminant hepatitis B. *J Med Virol*, 65(3):485–92.

Pachter, L., Alexandersson, M., and Cawley, S. (2002). Applications of generalized pair hidden Markov models to alignment and gene finding problems. *J Comput Biol*, 9(2):389–99.

Papatheodoridis, G. V., Manesis, E., and Hadziyannis, S. J. (2001). The long-term outcome of interferon-alpha treated and untreated patients with HBeAg-negative chronic hepatitis B. *J Hepatol*, 34(2):306–13.

Parker, L. T., Zakeri, H., Deng, Q., Spurgeon, S., Kwok, P. Y., and Nickerson, D. A. (1996). AmpliTaq DNA polymerase, FS dye-terminator sequencing: analysis of peak height patterns. *Biotechniques*, 21(4):694–9.

Patient, R., Hourioux, C., Sizaret, P. Y., Trassard, S., Sureau, C., and Roingeard, P. (2007). Hepatitis B virus subviral envelope particle morphogenesis and intracellular trafficking. *J Virol*, 81(8):3842–51.

Peng, Z. G., Zhao, Z. Y., Li, Y. P., Wang, Y. P., Hao, L. H., Fan, B., Li, Y. H., Wang, Y. M., Shan, Y. Q., Han, Y. X., Zhu, Y. P., Li, J. R., You, X. F., Li, Z. R., and Jiang, J. D. (2011). Host apolipoprotein B messenger RNA-editing enzyme catalytic polypeptide-like 3G is an innate defensive factor and drug target against hepatitis C virus. *Hepatology*, 53(4):1080–9.

Perlman, D. H., Berg, E. A., O'connor, P. B., Costello, C. E., and Hu, J. (2005). Reverse transcription-associated dephosphorylation of hepadnavirus nucleocapsids. *Proc Natl Acad Sci USA*, 102(25):9020–5.

Prange, R. and Streeck, R. E. (1995). Novel transmembrane topology of the hepatitis B virus envelope proteins. *EMBO J*, 14(2):247–56.

Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., Cocuzza, A. J., Jensen, M. A., and Baumeister, K. (1987). A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science*, 238(4825):336–41.

Purdy, J. B., Gafni, R. I., Reynolds, J. C., Zeichner, S., and Hazra, R. (2008). Decreased bone mineral density with off-label use of tenofovir in children and adolescents infected with human immunodeficiency virus. *J Pediatr*, 152(4):582–4.

Purdy, M. A., Talekar, G., Swenson, P., Araujo, A., and Fields, H. (2007). A new algorithm for deduction of hepatitis B surface antigen subtype determinants from the amino acid sequence. *Intervirology*, 50(1):45–51.

Qiu, P., Soder, G. J., Sanfiorenzo, V. J., Wang, L., Greene, J. R., Fritz, M. A., and Cai, X. Y. (2003). Quantification of single nucleotide polymorphisms by automated DNA sequencing. *Biochem Biophys Res Commun*, 309(2):331–8.

Rabe, B., Glebe, D., and Kann, M. (2006). Lipid-mediated introduction of hepatitis B virus capsids into nonsusceptible cells allows highly efficient replication and facilitates the study of early infection events. *J Virol*, 80(11):5465–73.

Ramsey, F. P. (1931). *Truth and Probability*. Harcourt, Brace and Company, New York.

Rapti, I., Dimou, E., Mitsoula, P., and Hadziyannis, S. J. (2007). Adding-on versus switching-to adefovir therapy in lamivudine-resistant HBeAg-negative chronic hepatitis B. *Hepatology*, 45(2):307–13.

Reeve, M. A. and Fuller, C. W. (1995). A novel thermostable polymerase for DNA sequencing. *Nature*, 376(6543):796–7.

Rijckborst, V., Hansen, B. E., Cakaloglu, Y., Ferenci, P., Tabak, F., Akdogan, M., Simon, K., Akarca, U. S., Flisiak, R., Verhey, E., Van Vuuren, A. J., Boucher, C. A., ter Borg, M. J., and Janssen, H. L. (2010). Early on-treatment prediction of response to peginterferon alfa-2a for HBeAg-negative chronic hepatitis B using HBsAg and HBV DNA levels. *Hepatology*, 52(2):454–61.

Rijckborst, V., Hansen, B. E., Ferenci, P., Brunetto, M. R., Tabak, F., Cakaloglu, Y., Lanza, A. G., Messina, V., Iannacone, C., Massetto, B., Regep, L., Colombo, M., Janssen, H. L., and Lampertico, P. (2012). Validation of a stopping rule at week 12 using HBsAg and HBV DNA for HBeAg-negative patients treated with peginterferon alfa-2a. *J Hepatol*, 56(5):1006–11.

Rogozin, I. B., Basu, M. K., Jordan, I. K., Pavlov, Y. I., and Koonin, E. V. (2005). APOBEC4, a new member of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases predicted by computational analysis. *Cell Cycle*, 4(9):1281–5.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem*, 242(1):84–9.

Ronaghi, M., Uhlén, M., and Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. *Science*, 281(5375):363, 365.

Rose, P. P. and Korber, B. T. (2000). Detecting hypermutations in viral sequences with an emphasis on G –> A hypermutation. *Bioinformatics*, 16(4):400–1.

Rosenblum, B. B., Lee, L. G., Spurgeon, S. L., Khan, S. H., Menchen, S. M., Heiner, C. R., and Chen, S. M. (1997). New dye-labeled terminators for improved DNA sequencing patterns. *Nucleic Acids Res*, 25(22):4500–4.

Rösler, C., Köck, J., Kann, M., Malim, M. H., Blum, H. E., Baumert, T. F., and von Weizsäcker, F. (2005). APOBEC-mediated interference with hepadnavirus production. *Hepatology*, 42(2):301–9.

Rozanov, M., Plikat, U., Chappey, C., Kochergin, A., and Tatusova, T. (2004). A web-based genotyping resource for viral sequences. *Nucleic Acids Res*, 32(Web Server issue):W654–9.

Sánchez-Tapias, J. M., Costa, J., Mas, A., Bruguera, M., and Rodés, J. (2002). Influence of hepatitis B virus genotype on the long-term outcome of chronic hepatitis B in western patients. *Gastroenterology*, 123(6):1848–56.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M., and Smith, M. (1977b). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596):687–95.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977a). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*, 74(12):5463–7.

Savage, L. (1961). *The Subjective Basis of Statistical Practice*. University of Michigan.

Savenko, V. (2013). Genotyping complex viral dual infections involving recombinant forms using population-based sequence data. Bachelor's thesis, Saarland University.

Scaglioni, P. P., Melegari, M., and Wands, J. R. (1997). Biologic properties of hepatitis B viral genomes with mutations in the precore promoter and precore open reading frame. *Virology*, 233(2):374–81.

Schädler, S. and Hildt, E. (2009). HBV life cycle: entry and morphogenesis. *Viruses*, 1(2):185–209.

Schaefer, S. (2007). Hepatitis B virus taxonomy and hepatitis B virus genotypes. *World J Gastroenterol*, 13(1):14–21.

Schildgen, O., Schewe, C. K., Vogel, M., Däumer, M., Kaiser, R., Weitner, L., Matz, B., and Rockstroh, J. K. (2004). Successful therapy of hepatitis B with tenofovir in hiv-infected patients failing previous adefovir and lamivudine treatment. *AIDS*, 18(17):2325–7.

Schröfelbauer, B., Yu, Q., Zeitlin, S. G., and Landau, N. R. (2005). Human immunodeficiency virus type 1 Vpr induces the degradation of the UNG and SMUG uracil-DNA glycosylases. *J Virol*, 79(17):10978–87.

Schultz, A. K., Bulla, I., Abdou-Chekaraou, M., Gordien, E., Morgenstern, B., Zoaulim, F., Dény, P., and Stanke, M. (2012). jpHMM: recombination analysis in viruses with circular genomes such as the hepatitis B virus. *Nucleic Acids Res*, 40(Web Server issue):W193–8.

Schultz, A. K., Zhang, M., Bulla, I., Leitner, T., Korber, B., Morgenstern, B., and Stanke, M. (2009). jpHMM: improving the reliability of recombination prediction in HIV-1. *Nucleic Acids Res*, 37(Web Server issue):W647–51.

Schultz, A. K., Zhang, M., Leitner, T., Kuiken, C., Korber, B., Morgenstern, B., and Stanke, M. (2006). A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. *BMC Bioinformatics*, 7:265.

Schulze, A., Gripon, P., and Urban, S. (2007). Hepatitis B virus infection initiates with a large surface protein-dependent binding to heparan sulfate proteoglycans. *Hepatology*, 46(6):1759–68.

Seroussi, Y. and Seroussi, E. (2007). TraceHaplotyper: using direct sequencing to determine the phase of an indel followed by biallelic SNPs. *Biotechniques*, 43(4):452, 454, 456.

Sheehy, A. M., Gaddis, N. C., Choi, J. D., and Malim, M. H. (2002). Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature*, 418(6898):646–50.

Short, J. M., Chen, S., Roseman, A. M., Butler, P. J., and Crowther, R. A. (2009). Structure of hepatitis B surface antigen from subviral tubes determined by electron cryomicroscopy. *J Mol Biol*, 390(1):135–41.

Simmonds, P. (2001). Reconstructing the origins of human hepatitis viruses. *Philos Trans R Soc Lond B Biol Sci*, 356(1411):1013–26.

Simmonds, P. and Midgley, S. (2005). Recombination in the genesis and evolution of hepatitis B virus genotypes. *J Virol*, 79(24):15467–76.

Smith, D. M., Richman, D. D., and Little, S. J. (2005). Hiv superinfection. *J Infect Dis*, 192(3):438–44.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B., and Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–9.

Sodeik, B. (2000). Mechanisms of viral transport in the cytoplasm. *Trends Microbiol*, 8(10):465–72.

Sonneveld, M. J., Hansen, B. E., Piratvisuth, T., Jia, J. D., Zeuzem, S., Gane, E., Liaw, Y. F., Xie, Q., Heathcote, E. J., Chan, H. L., and Janssen, H. L. (2013). Response-guided peginterferon therapy in hepatitis B e antigen-positive chronic hepatitis B using serum hepatitis B surface antigen levels. *Hepatology*, 58(3):872–80.

Sonneveld, M. J., Rijckborst, V., Boucher, C. A., Hansen, B. E., and Janssen, H. L. (2010). Prediction of sustained response to peginterferon alfa-2b for hepatitis B e antigen-positive chronic hepatitis B using on-treatment hepatitis B surface antigen decline. *Hepatology*, 52(4):1251–7.

Sousa Santos, C., Robalo, J. I., Collares-Pereira, M. J., and Almada, V. C. (2005). Heterozygous indels as useful tools in the reconstruction of DNA sequences and in the assessment of ploidy level and genomic constitution of hybrid organisms. *DNA Seq*, 16(6):462–7.

Sova, P. and Volsky, D. J. (1993). Efficiency of viral DNA synthesis during infection of permissive and nonpermissive cells with vif-negative human immunodeficiency virus type 1. *J Virol*, 67(10):6322–6.

Struck, D., Perez-Bercoff, D., Devaux, C., and Schmit, J. C. (2010). COMET: a novel approach to HIV-1 subtype prediction. In *8th European HIV Drug Resistance Workshop, Sorrento, Italy*.

Stuyver, L., De Gendt, S., Van Geyt, C., Zoulim, F., Fried, M., Schinazi, R. F., and Rossau, R. (2000). A new genotype of hepatitis B virus: complete genome and phylogenetic relatedness. *J Gen Virol*, 81(Pt 1):67–74.

Stuyver, L. J., Locarnini, S. A., Lok, A., Richman, D. D., Carman, W. F., Dienstag, J. L., and Schinazi, R. F. (2001). Nomenclature for antiviral-resistant human hepatitis B virus mutations in the polymerase region. *Hepatology*, 33(3):751–7.

Summers, J. and Mason, W. S. (1982). Replication of the genome of a hepatitis B–like virus by reverse transcription of an RNA intermediate. *Cell*, 29(2):403–15.

Suspène, R., Aynaud, M. M., Guétard, D., Henry, M., Eckhoff, G., Marchio, A., Pineau, P., Dejean, A., Vartanian, J. P., and Wain-Hobson, S. (2011). Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. *Proc Natl Acad Sci USA*, 108(12):4858–63.

Suspène, R., Guétard, D., Henry, M., Sommer, P., Wain-Hobson, S., and Vartanian, J. P. (2005b). Extensive editing of both hepatitis B virus DNA strands by APOBEC3 cytidine deaminases in vitro and in vivo. *Proc Natl Acad Sci USA*, 102(23):8321–6.

Suspène, R., Henry, M., Guillot, S., Wain-Hobson, S., and Vartanian, J. P. (2005a). Recovery of APOBEC3-edited human immunodeficiency virus G->A hypermutants by differential DNA denaturation PCR. *J Gen Virol*, 86(Pt 1):125–9.

Suzuki, Y., Kobayashi, M., Ikeda, K., Suzuki, F., Arfase, Y., Akuta, N., Hosaka, T., Saitoh, S., Someya, T., Matsuda, M., Sato, J., Watabiki, S., Miyakawa, Y., and Kumada, H. (2005). Persistence of acute infection with hepatitis B virus genotype A and treatment in Japan. *J Med Virol*, 76(1):33–9.

Swerdlow, H., Wu, S. L., Harke, H., and Dovichi, N. J. (1990). Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *J Chromatogr*, 516(1):61–7.

Szmuness, W., Stevens, C. E., Harley, E. J., Zang, E. A., Oleszko, W. R., William, D. C., Sadovsky, R., Morrison, J. M., and Kellner, A. (1980). Hepatitis B vaccine: demonstration of efficacy in a controlled clinical trial in a high-risk population in the United States. *N Engl J Med*, 303(15):833–41.

Tabor, S. and Richardson, C. C. (1989). Selective inactivation of the exonuclease activity of bacteriophage T7 DNA polymerase by in vitro mutagenesis. *J Biol Chem*, 264(11):6447–58.

Tabor, S. and Richardson, C. C. (1990). DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. Effect of pyrophosphorolysis and metal ions. *J Biol Chem*, 265(14):8322–8.

Tabor, S. and Richardson, C. C. (1995). A single residue in DNA polymerases of the Escherichia coli DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc Natl Acad Sci USA*, 92(14):6339–43.

Takada, S. and Koike, K. (1994). Three sites of the hepatitis B virus X protein cooperatively interact with cellular proteins. *Virology*, 205(2):503–10.

Tallo, T., Tefanova, V., Priimägi, L., Schmidt, J., Katargina, O., Michailov, M., Mukomolov, S., Magnius, L., and Norder, H. (2008). D2: major subgenotype of hepatitis B virus in Russia and the Baltic region. *J Gen Virol*, 89(Pt 8):1829–39.

Tanaka, Y., Marusawa, H., Seno, H., Matsumoto, Y., Ueda, Y., Kodama, Y., Endo, Y., Yamauchi, J., Matsumoto, T., Takaori-Kondo, A., Ikai, I., and Chiba, T. (2006). Anti-viral protein APOBEC3G is induced by interferon-alpha stimulation in human hepatocytes. *Biochem Biophys Res Commun*, 341(2):314–9.

Tatematsu, K., Tanaka, Y., Kurbanov, F., Sugauchi, F., Mano, S., Maeshiro, T., Nakayoshi, T., Wakuta, M., Miyakawa, Y., and Mizokami, M. (2009). A genetic variant of hepatitis B virus divergent from known human and ape genotypes isolated from a Japanese patient and provisionally assigned to new genotype J. *J Virol*, 83(20):10538–47.

Teng, B., Burant, C. F., and Davidson, N. O. (1993). Molecular cloning of an apolipoprotein B messenger RNA editing protein. *Science*, 260(5115):1816–9.

Tenney, D. J., Rose, R. E., Baldick, C. J., Pokornowski, K. A., Eggers, B. J., Fang, J., Wichroski, M. J., Xu, D., Yang, J., Wilber, R. B., and Colonno, R. J. (2009). Long-term monitoring shows hepatitis B virus resistance to entecavir in nucleoside-naïve patients is rare through 5 years of therapy. *Hepatology*, 49(5):1503–14.

Thakur, V., Guptan, R. C., Kazim, S. N., Malhotra, V., and Sarin, S. K. (2002). Profile, spectrum and significance of HBV genotypes in chronic liver disease patients in the Indian subcontinent. *J Gastroenterol Hepatol*, 17(2):165–70.

Tran, T. T., Trinh, T. N., and Abe, K. (2008). New complex recombinant genotype of hepatitis B virus identified in Vietnam. *J Virol*, 82(11):5657–63.

Turelli, P., Mangeat, B., Jost, S., Vianin, S., and Trono, D. (2004). Inhibition of hepatitis B virus replication by APOBEC3G. *Science*, 303(5665):1829.

Usuda, S., Okamoto, H., Iwanari, H., Baba, K., Tsuda, F., Miyakawa, Y., and Mayumi, M. (1999). Serological detection of hepatitis B virus genotypes by ELISA with monoclonal antibodies to type-specific epitopes in the preS2-region product. *J Virol Methods*, 80(1):97–112.

van Zonneveld, M., Honkoop, P., Hansen, B. E., Niesters, H. G., Darwish Murad, S., de Man, R. A., Schalm, S. W., and Janssen, H. L. (2004). Long-term follow-up of alpha-interferon treatment of patients with chronic hepatitis B. *Hepatology*, 39(3):804–10.

Vapnik, V. (1996). *The Nature of Statistical Learning Theory*. Springer.

Vartanian, J. P., Henry, M., Marchio, A., Suspène, R., Aynaud, M. M., Guétard, D., Cervantes-Gonzalez, M., Battiston, C., Mazzaferro, V., Pineau, P., Dejean, A., and Wain-Hobson, S. (2010). Massive APOBEC3 editing of hepatitis B viral DNA in cirrhosis. *PLoS Pathog*, 6(5):e1000928.

Vartanian, J. P., Meyerhans, A., Asjö, B., and Wain-Hobson, S. (1991). Selection, recombination, and G--->A hypermutation of human immunodeficiency virus type 1 genomes. *J Virol*, 65(4):1779–88.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz,

S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. (2001). The sequence of the human genome. *Science*, 291(5507):1304–51.

Vieira, V. C. and Soares, M. A. (2013). The role of cytidine deaminases on innate immune responses against human viral infections. *Biomed Res Int*, 2013:683095.

Visvanathan, K., Skinner, N. A., Thompson, A. J., Riordan, S. M., Sozzi, V., Edwards, R., Rodgers, S., Kurtovic, J., Chang, J., Lewin, S., Desmond, P., and Locarnini, S. (2007). Regulation of Toll-like receptor-2 expression in chronic hepatitis B by the precore protein. *Hepatology*, 45(1):102–10.

Viviani, S., Jack, A., Hall, A. J., Maine, N., Mendy, M., Montesano, R., and Whittle, H. C. (1999). Hepatitis B vaccination in infancy in The Gambia: protection against carriage at 9 years of age. *Vaccine*, 17(23-24):2946–50.

von Schwedler, U., Song, J., Aiken, C., and Trono, D. (1993). Vif is crucial for human immunodeficiency virus type 1 proviral DNA synthesis in infected cells. *J Virol*, 67(8):4945–55.

Wain-Hobson, S., Sonigo, P., Guyader, M., Gazit, A., and Henry, M. (1995). Erratic G-->A hypermutation within a complete caprine arthritis-encephalitis virus (CAEV) provirus. *Virology*, 209(2):297–303.

Walsh, R. and Locarnini, S. (2012). Hepatitis B precore protein: pathogenic potential and therapeutic promise. *Yonsei Med J*, 53(5):875–85.

Wang, G. H. and Seeger, C. (1992). The reverse transcriptase of hepatitis B virus acts as a protein primer for viral DNA synthesis. *Cell*, 71(4):663–70.

Wang, L. (2005b). *Support Vector Machines: Theory and Applications*, volume 177 of *Studies in Fuzziness and Soft Computing*. Springer Berlin, Heidelberg, Germany.

Wang, Z., Liu, Z., Zeng, G., Wen, S., Qi, Y., Ma, S., Naoumov, N. V., and Hou, J. (2005a). A new intertype recombinant between genotypes C and D of hepatitis B virus identified in China. *J Gen Virol*, 86(Pt 4):985–90.

Weber, M., Bronsema, V., Bartos, H., Bosserhoff, A., Bartenschlager, R., and Schaller, H. (1994). Hepadnavirus P protein utilizes a tyrosine residue in the TP domain to prime reverse transcription. *J Virol*, 68(5):2994–9.

Wedekind, J. E., Dance, G. S., Sowden, M. P., and Smith, H. C. (2003). Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. *Trends Genet*, 19(4):207–16.

Wei, Y., Tavis, J. E., and Ganem, D. (1996). Relationship between viral DNA synthesis and virion envelopment in hepatitis B viruses. *J Virol*, 70(9):6455–8.

WHO (2013). Hepatitis B (fact sheet 204).

Won, K. J., Hamelryck, T., Prügel-Bennett, A., and Krogh, A. (2007). An evolutionary method for learning HMM structure: prediction of protein secondary structure. *BMC Bioinformatics*, 8:357.

Wynne, S. A., Crowther, R. A., and Leslie, A. G. (1999). The crystal structure of the human hepatitis B virus capsid. *Mol Cell*, 3(6):771–80.

Yan, H., Zhong, G., Xu, G., He, W., Jing, Z., Gao, Z., Huang, Y., Qi, Y., Peng, B., Wang, H., Fu, L., Song, M., Chen, P., Gao, W., Ren, B., Sun, Y., Cai, T., Feng, X., Sui, J., and Li, W. (2012). Sodium taurocholate cotransporting polypeptide is a functional receptor for human hepatitis B and D virus. *Elife*, 1:e00049.

Yang, J., Xing, K., Deng, R., Wang, J., and Wang, X. (2006). Identification of Hepatitis B virus putative intergenotype recombinants by using fragment typing. *J Gen Virol*, 87(Pt 8):2203–15.

Yang, W. and Summers, J. (1995). Illegitimate replication of linear hepadnavirus DNA through nonhomologous recombination. *J Virol*, 69(7):4029–36.

Yim, H. J. and Lok, A. S. (2006). Natural history of chronic hepatitis B virus infection: what we knew in 1981 and what we know in 2005. *Hepatology*, 43(2 Suppl 1):S173–81.

Yoon, B. J. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr Genomics*, 10(6):402–15.

Yu, H., Yuan, Q., Ge, S. X., Wang, H. Y., Zhang, Y. L., Chen, Q. R., Zhang, J., Chen, P. J., and Xia, N. S. (2010). Molecular and phylogenetic analyses suggest an additional hepatitis B virus genotype "I". *PLoS One*, 5(2):e9297.

Yu, Q., Chen, D., König, R., Mariani, R., Unutmaz, D., and Landau, N. R. (2004). APOBEC3B and APOBEC3C are potent inhibitors of simian immunodeficiency virus replication. *J Biol Chem*, 279(51):53379–86.

Yuen, M. F., Tanaka, Y., Mizokami, M., Yuen, J. C., Wong, D. K., Yuan, H. J., Sum, S. M., Chan, A. O., Wong, B. C., and Lai, C. L. (2004). Role of hepatitis B virus genotypes Ba and C, core promoter and precore mutations on hepatocellular carcinoma: a case control study. *Carcinogenesis*, 25(9):1593–8.

Zhang, H. W., Yin, J. H., Li, Y. T., Li, C. Z., Ren, H., Gu, C. Y., Wu, H. Y., Liang, X. S., Zhang, P., Zhao, J. F., Tan, X. J., Lu, W., Schaefer, S., and Cao, G. W. (2008). Risk factors for acute hepatitis B and its progression to chronic hepatitis in Shanghai, China. *Gut*, 57(12):1713–20.

Zhang, M., Schultz, A. K., Calef, C., Kuiken, C., Leitner, T., Korber, B., Morgenstern, B., and Stanke, M. (2006). jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1. *Nucleic Acids Res*, 34(Web Server issue):W463–5.

Zhang, Q., Wu, G., Richards, E., Jia, S., and Zeng, C. (2007). Universal primers for HBV genome DNA amplification across subtypes: a case study for designing more effective viral primers. *Virol J*, 4:92.

Zlotnick, A., Cheng, N., Stahl, S. J., Conway, J. F., Steven, A. C., and Wingfield, P. T. (1997). Localization of the C terminus of the assembly domain of hepatitis B virus capsid protein: implications for morphogenesis and organization of encapsidated RNA. *Proc Natl Acad Sci USA*, 94(18):9556–61.

Zoulim, F., Durantel, D., and Deny, P. (2009). Management and prevention of drug resistance in chronic hepatitis B. *Liver Int*, 29 Suppl 1:108–15.

Zoulim, F., Saputelli, J., and Seeger, C. (1994). Woodchuck hepatitis virus X protein is required for viral infection in vivo. *J Virol*, 68(3):2026–30.

List of Own Publications

Bastian Beggel, Maria Neumann-Fraune, Rolf Kaiser, Jens Verheyen, Thomas Lengauer (2013). Inferring short-range linkage information from sequencing chromatograms. *PLoS ONE*, 8(12): e81687. doi:10.1371/journal.pone.0081687.

Bastian Beggel, Carsten Münk, Martin Dämer, Katharina Hauck, Dieter Hässinger, Thomas Lengauer, Andreas Erhardt (2013). Full genome ultra-deep pyrosequencing associates G-to-A hypermutation of the hepatitis B virus genome with the natural progression of hepatitis B. *Journal of Viral Hepatitis*, 20(12):882-9.

Maria Neumann-Fraune, **Bastian Beggel**, Herbert Pfister, Rolf Kaiser, Jens Verheyen (2013). High frequency of complex mutational patterns in lamivudine resistant hepatitis B virus isolates. *Journal of Medical Virology*, 85(5):775-9.

Bastian Beggel, Maria Neumann-Fraune, Matthias Döring, Glenn Lawyer, Rolf Kaiser, Jens Verheyen, Thomas Lengauer (2012). Genotyping hepatitis B virus dual infections using population-based sequence data. *Journal of General Virology*, 93, 1899Ű1907.

Valentina Svichera, Valeria Cento, Romina Salpini, Fabio Mercurio, Maria Fraune, **Bastian Beggel**, Yue Han, Caterina Gori, Linda Wittkop, Ada Bertoli, Valeria Micheli, Guido Gubertini, Roberta Longo, Sara Romano, Michela Visca, Valentina Gallinaro, Nicoletta Marino, Francesco Mazzotta, Giuseppe Maria De Sanctis, Hervè Fleury, Pascale Trimoulet, Mario Angelico, Giuseppina Cappiello, Xin Xin Zhang, Jens Verheyen, Francesca Ceccherini-Silberstein, Carlo Federico Perno (2011). Role of hepatitis B virus genetic barrier in drug-resistance and immune-escape development. *Digestive and Liver Disease*, 43(12):975-83.

Stefan Reuter, Mark Oette, Frank Clemens Wilhelm, **Bastian Beggel**, Rolf Kaiser, Melanie Balduin, Finja Schweitzer, Jens Verheyen, Ortwin Adams, Thomas Lengauer, Gerd Fätkenheuer, Herbert Pfister, Dieter Häussinger (2011). Prevalence and characteristics of hepatitis B and C virus infections in treatment-naïve HIV-infected patients. *Medical Microbiology and Immunology*, 200(1):39-49.

Stefan Reuter, Mark Oette, Frank Clemens Wilhelm, **Bastian Beggel**, Rolf Kaiser, Melanie Balduin, Finja Schweitzer, Jens Verheyen, Ortwin Adams, Thomas Lengauer, Gerd Fätkenheuer, Herbert Pfister, Dieter Häussinger (2011). Erratum to: Prevalence and characteristics of hepatitis B and C virus infections in treatment-naïve HIV-infected patients. *Medical Microbiology and Immunology*, 200(1):51.

Soo-Yon Rhee, Severine Margeridon-Thermet, Mindie H. Nguyen, Tommy F. Liua, Ron M. Kagan, **Bastian Beggel**, Jens Verheyen, Rolf Kaiser, Robert W. Shafer (2010). Hepatitis B virus reverse transcriptase sequence variant database for sequence analysis and mutation discovery. *Antiviral Research*, 88(3):269-75.